Methodology Advisory Committee

18 June 2004

Statistical Matching of the HES and NHS: An Exploration of Issues in the use of Unconstrained and Constrained Approaches in Creating a Basefile for a Microsimulation Model of the Pharmaceutical Benefits Scheme

Technical Working Group (TWG)
ABS-NATSEM Collaboration on Statistical Matching

For comments or questions, contact:
Ruel Abello of the ABS (ruel.abello@abs.gov.au) or
Ben Phillips of the National Centre for Social and Economic Modelling (NATSEM) (Ben.Phillips@torrens.natsem.canberra.edu.au)

Other members of the TWG include Daniel Elazar, Guia Alcausin and Chris Gordon of the ABS, and Annie Abello, Laurie Brown and Sharyn Lymer of the NATSEM.

Executive Summary

Statistical matching is a procedure used to link two files or datasets where each record from one of the files is matched with a record from the second file that generally does not represent the same unit, but does represent a *similar* unit.

The constrained and unconstrained approaches to statistical matching are investigated in this paper. The issues associated with these approaches are identified and discussed. The conditional independence assumption, for example, is inherent in the procedure. Its implication for the analysis to be done using the matched dataset must be considered carefully.

While unconstrained matching gives the closest possible match between similar pairs, constrained matching has the advantage of replicating the marginal distributions in the donor file.

These traditional approaches to statistical matching are used to match two ABS datasets: the 1998-99 Household Expenditure Survey (HES) and the 2001 National Health Survey (NHS). The matching was done to explore building a base dataset for a microsimulation model of the Pharmaceutical Benefits Scheme (PBS). The main objective was to replicate the family structures of HES into the NHS.

Constrained matching, using linear programming, was found to be a better approach in synthetically creating completely enumerated families, and making sure that persons on the NHS are sensibly assigned to families using the HES family structure.

This paper is a preliminary output from a Technical Working Group comprising MD staff and the National Centre for Social and Economic Modelling (NATSEM). The former's main interest is to explore methodological issues associated with statistical matching procedures. The latter developed the microsimulation model of the PBS and relies on ABS microdatasets to create base files for the said model. It has done the preliminary statistical matching reported in this paper.

Questions for MAC members

1. Have we correctly applied the conventional methods of statistical matching?
2. Have we correctly identified all the relevant issues associated with the use of these methods?
3. What are the implications of the conditional independence assumption-- an assumption inherent in statistical matching?
4. How accurate or valid must statistical matches be? For what purposes and under what conditions are the results of the statistical matching sufficiently accurate or valid?
5. What types of sensitivity analyses are appropriate to check the robustness of results from various statistical matching methods?
6. As applied in this paper, is statistical matching a suitable procedure to create synthetic estimates of family structure?
7. What are the most promising alternative methods to pursue?

Statistical Matching of the HES and NHS Survey Files: An Exploration of Issues in the Use of Unconstrained and Constrained Approaches in Creating A Basefile for a Microsimulation Model of the Pharmaceutical Benefits Scheme

1 Introduction

1 Matching, or record linkage, is the process by which records or units from different sources are combined into a single file. This is done primarily to create a composite dataset that augments the variables in one dataset with the variables available in another.

2 Combining datasets increases the power of analysts to understand socioeconomic phenomena. In the field of microsimulation, for example, "what if" analyses employ information about individuals or families contained in survey microdata. Understandably, a single survey microdata will not always contain all the variables required by an analyst. But many separate survey microdata have variables or identifiers that are common to each other. This makes linking the different datasets, based on these common variables, possible. The resulting linked or matched dataset allows a more comprehensive modeling of a phenomenon in question.

3 Producing an augmented unit record data, where additional variables from a donor data are added to a base data, is a primary objective of matching. There are applications however where *variable construction*, rather than *variable addition*, is the objective of matching. In the practical application discussed in Chapter X of this paper, we shall see that the preliminary reason for matching the two datasets is to create an additional synthetic variable in one dataset that mimics the family structure of units in another, and not necessarily to use the complete set of variables in the matched data.

1.1 Types of matching

4 There are two main types of matching: exact matching and statistical matching. In exact matching, information about a particular record on one file is linked to information on another file, thus creating a single file with expanded information at the level of the record. The linking is done using identifiers which allows information about the *same* individual unit to be identified in the two files.

5 However, it is often the case that identifiers are not available in one or both files, or that there is little or no overlap between the records in the two files- as in the case when each of these two files consists of survey sample of a large population. In cases like this, statistical matching is used, where each record from one of the files is matched with a record from the second file that generally does not represent the same unit, but does represent a *similar* unit (Rodgers 1984 p. 91).

6 The focus of this paper is statistical matching, and in particular, the issues associated with the use of unconstrained and constrained approaches. These two approaches are known in the literature as the 'traditional' or 'conventional' approaches to statistical matching. In this paper, we describe their use in matching two ABS surveys, the 2001 National Health Survey and the 1998-99 Household Expenditure Survey. These two datasets were matched for the purpose of replicating the family structures evident in HES in a matched data that resembles the marginal distributions of the NHS.

1.2 Outline

7 The paper gives an overview of statistical matching; explains the methods involved in the conventional approaches of statistical matching; explores the major assumption inherent in these methods and how this assumption may impinge on any analysis using the statistically matched data; explores how one can assess the validity of matched data; demonstrates the use of these conventional methods using two ABS survey microdata, the 1998 Household Expenditure survey (HES) and the 2001 National Health Survey (NHS); explains why these two microdatasets are being matched (i.e. to create a base data file for a microsimulation model of the Pharmaceutical Benefits Scheme (PBS); assesses the validity of the resulting matched data; and then draws some conclusions and identifies other approaches for matching that may be done in the future.

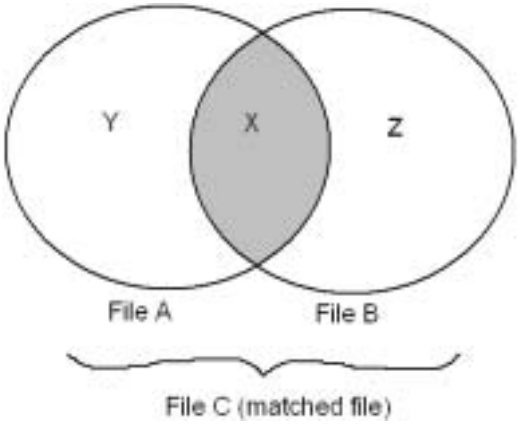
2 Statistical matching: an overview

8 Rodgers (1984) gives a clear description of statistical matching. (Other materials that have good overviews include Rassler (2002), Radner et al (1980), Cohen (1991), and Moriarty and Scheuren (2001)).

9 Statistical matching can be used to link two or more files, but for purposes of brevity, the method is described here in terms of two files only, called File A (or the base file) and File B (the donor file), taken from two different surveys.

10 File A contains a vector of variables X and Y, while file B contains a vector of variables X and Z. X therefore are variables that are common to both files A and B; Y are variables unique to File A, while Z are variables unique to File B. The purpose of statistical matching is to generate-- by combining Files A and B-- a composite file containing (X,Y,Z).

Figure 1. Statistical Matching



11 As mentioned earlier, if the files to be combined come from two different surveys with no overlap of the samples at all, then with statistical matching the records being combined are for *similar* entities, and not for the *same* entities. The X variables are used as "matching" variables, to create the matched file.

2.1 Statistical matching: conventional approaches

12 Conventional statistical matching methods (or traditional methods, as some authors call them) can be grouped broadly into two types: unconstrained and constrained. To explain the difference between these two methods, we borrow Rodgers' (1984, pp 91-93) notations and examples, as follows.

13 Let there be a vector **S** of variables for each of n_A records on file A, and let there be a vector **T** of variables for each of n_B records on file. Both vectors consist of indicators of a common set of characteristics of the analysis units, say persons, and we refer to these as the **X** variables, **X** = (X_1, ..., X_P). The remaining variables in file A are referred to as **Y** variables, where **Y** = ((Y_1, ..., Y_Q). The remaining variables in File B are referred to as the **Z** variables**,** where **Z** = ((Z_1, ..., Z_R). Also, a sampling weight, **w**, may be associated with each record in both files (Rodgers 1984, p 91-92).

14 The following highly simplified example of records from Files A and B from Rodgers (1984, p. 92) will be used to highlight the difference between unconstrained and

constrained statistical matching. Table 1 shows the records in File A, while Table 1 shows the records in File B.

Table 1. Simplified example of a File A

Case	Sex x_1^A	Age x_2^A	Expenditure y^A	Weight w_i^A
A1	M	42	9.156	3
A2	M	35	9.149	3
A3	F	63	9.287	3
A4	M	55	9.512	3
A5	F	28	8.494	3
A6	F	53	8.891	3
A7	F	22	8.425	3
A8	M	25	8.867	3
Mean	.50	40.38	8.97	
SD	.53	15.32	0.38	

Source: Table 1a in Rodgers (1984).

Table 2. Simplified example of a File B

Case	Sex x_1^B	Age x_2^B	Income z^B	Weight w_j^B
B1	F	33	6.935	4
B2	M	52	5.524	4
B3	M	28	4.223	4
B4	F	59	6.147	4
B5	M	41	7.243	4
B6	F	45	3.230	4
Mean	.50	43.00	5.55	
SD	.55	11.58	1.57	

Source: Table 1b in Rodgers (1984).

15 An important element of statistical matching is the *distance function*, indicated by d_{ij} and is used to assess the similarity of any pair of cases based on the common variable X.

16 A distance function is defined as the absolute differences in the X variable of two cases: $d_{ij} = | x_i - x_j |$. If there are more than one X variable to be used in assessing the similarity of pairs, then these X variables must assume some weights in the distance function, for example, the weights a and b in $d_{ij} = a|x_{1_i} - x_{1_j}| + b|x_{2_i} - x_{2_j}|$. These weights can be assigned subjectively, or can be estimated from other regressions or factor analyses.

17 In statistical matching, a weight is assigned to each record in the matched file, called w_{ij}, which may be may be equal to w_i-- the weight associated with the input case from file A-- or it may be modified depending on the matching technique and the need to align units in the files being matched (Rodgers 1984, p. 92).

2.1.1 Unconstrained matching

18 A statistical match is said to be unconstrained if there are no restrictions on the number of A records to which the values of the Z variables in file B can be imputed. Using Rodger's example, this is illustrated in Table 3 below where-- using sex as a matching group-- each person in file A is matched with the person in file B whose age is closest to his/her own. Unconstrained matching is sometimes referred to as matching **'with replacement'**, evident in Table 3 where the same unique person in File B can be matched with several persons in File A. For example, B5 has been matched to both A1 and A2. It is also possible that a file B record is not attached to *any* file A record (e.g. B6 did not get matched to any A record).

19 Unconstrained matching is a 'nearest-neighbour' matching technique: it allows for the closest possible match for each A record. But this comes at a cost, which is increasing the sample variance of estimators involving the **Z** variables. This is shown in the example by both the mean and standard deviation of the **Z** variable in the matched file, which have now differed from the corresponding statistics in file B (Rodgers 1984 p 92).

Table 3. Unconstrained match (example)

A Case	B Case	$x_1^A=x_1^B$	x_2^A	x_2^B	d_{ij}	y^A	z^B	w_{ij}
A1	**B5**	M	42	41	1	9.156	7.243	3
A2	**B5**	M	35	41	6	9.149	7.243	3
A3	**B4**	F	63	59	4	9.287	6.147	3
A4	**B2**	M	55	52	3	9.512	5.524	3
A5	**B1**	F	28	33	5	8.494	6.932	3
A6	**B4**	F	53	59	6	8.891	6.147	3
A7	**B1**	F	22	33	11	8.425	6.932	3
A8	**B3**	M	25	28	3	8.867	4.223	3
Mean		.50	40.38	43.25	4.88	8.97	6.30	
SD		.53	15.32	12.40	3.00	0.38	1.06	

Source: Table 1c in Rodgers (1984).

2.1.2 Constrained matching

20 An alternative to unconstrained matching is called **constrained** matching, also known as the linear programming (LP) method following the work of Barr and Turner (1978) and subsequently applied by many other authors.

21 Constrained matching requires the use of all records in Files A and B, and basically **preserves the marginal Y and Z distributions** (Barr and Turner 1978; Barr, Stewart and Turner 1982).

22 In a linear programming problem, the objective of constrained matching is to minimise the following function:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(d_{ij} * w_{ij}\right), \qquad (1)$$

where d_{ij} is the distance between cases i and j in files A and B, and w_{ij} is the weight to be allocated to records in the matched file, which are based on case *i* in file a and case *j* in file B, subject to the following conditions:

$$\sum_{j=1}^{m} w_{ij} = w_i, \quad \text{for } i = 1,...n \qquad (2)$$

and

$$\sum_{i=1}^{n} w_{ij} = w_j, \quad \text{for } j = 1,...m \qquad (3)$$

and

$$w_{ij} \geq 0 \text{ for all } i \text{ and } j. \qquad (4)$$

23 In Rodgers' example in Table 4, constrained matching makes use of all cases in both files (e.g. B6 is now matched to some A cases). The advantage of constrained matching is that the marginal distributions of Z as observed in File B are maintained in the matched file (Cohen 1991,p 65). Note that the mean and standard deviation of the Z variable are identical between File B and the matched file. The disadvantage of constrained matching is that the average distance between matched cases is greater than that obtained in unconstrained matching.

Table 4. Constrained match (example)

A Case	B Case	$x_1^A = x_1^B$	x_2^A	x_2^B	d_{ij}	Y^A	z^B	w_{ij}
A1	**B2**	M	42	52	10	9.156	5.524	1
A1	**B5**	M	42	41	1	9.156	7.243	2
A2	**B3**	M	35	28	7	9.149	4.223	1
A2	**B5**	M	35	41	6	9.149	7.243	2
A3	**B4**	F	63	59	4	9.287	6.147	3
A4	**B2**	M	55	52	3	9.512	5.524	3
A5	**B1**	F	28	33	5	8.494	6.932	3
A6	**B4**	F	53	59	6	8.891	6.147	1
A6	**B6**	F	53	45	8	8.891	3.230	2
A7	**B1**	F	22	33	11	8.425	6.932	1
A7	**B6**	F	22	45	23	8.425	3.230	2
A8	**B3**	M	25	28	3	8.867	4.223	3
	Mean	.50	40.38	43.00	6.46	8.97	5.55	
	SD	.53	15.32	11.58	5.81	0.38	1.57	

Source: Table 1d in Rodgers (1984).

2.2 Issues in constrained and unconstrained matching

2.2.1 Conditional independence assumption

24 An inherent assumption in statistical matching, whether constrained or unconstrained, is what is now popularly known as the *Conditional Independence Assumption* (CIA). It was originally pointed out by Sims (1972) and discussed extensively in subsequent literature. Statistical matching assumes that Y and Z, given X are independent. For the particular case of multivariate normal distributions of the variables, this is equivalent to the assumption that the partial correlations among the Y and Z variables, controlling on the X variables, are all zero (Rodgers 1984, p.93). Cohen 1991 (p74) writes: "Records from two files are matched or not matched on the basis of the values of X (in file A and file B). Therefore, there is no additional information in the matched file about the relationship between X and Y and between X and Z that is not explained by the relationships between X and Y and between X and Z."

25 Cohen (1991) shows a mathematical representation of the CIA in terms of the partial correlation between Y_i and Z_j conditioned on X, as:

$$\rho_{YZ.X} = \frac{\rho_{YZ} - \rho_{YX}\rho_{ZX}}{\left\{\left(1-\rho^2_{YX}\right)\left(1-\rho^2_{ZX}\right)\right\}^{1/2}} \qquad (5)$$

26 Using (5), Cohen explains that if both ρ_{YX} and ρ_{ZX} are close to 1, then the numerator of $\rho_{YZ.X}$ will be close to 0, or what amounts to the same thing, ρ_{YZ} will be close to 1. Cohen argues that to some extent this reasoning is valid, but noted how variable the correlation between Y and Z, ρ_{YZ}, can be even when ρ_{YX} and ρ_{ZX} are fairly close to 1 (Cohen 1991 p. 75). Cohen finds this variability "disturbing since the estimation of these correlations is presumably a major reason the statistical match was performed" (p. 75).

27 The CIA is a strong assumption and its *potential* seriousness has been pointed out by many authors, including Sims (1978), Rubin (1986), Paass (1986), and Rodgers and deVol (1982). Moriarty and Scheuren (2001) note that statistical matching is not a procedure to be applied casually because of this assumption. In the absence of auxiliary information (i.e. a third file containing auxiliary information about the full set X,Y,Z or the reduced set Y,Z), the authors argue, statistical matching is "unable to provide any sort of best estimate of the (Y,Z) relationship; the best that can be done is to exhibit variability for a range of plausible values of the (Y,Z) relationship". Singh et al (1993, p. 60) writes how important it is for the CIA to hold true: "The importance of the CIA is obvious, since the purpose of the match is to analyse the joint relationships of X, Y and Z. If the true relationships of the variables are such that conditional independence does not hold, then the CIA would mask an important component of these relationships, and would bias some analyses involving the full set of variables".

28 Despite the potential problems associated with CIA, statistical matching continues to be used, either using the constrained or unconstrained methods. This is because of the simplicity and flexibility of the methods, and of failure in many cases to prove that the CIA fails to hold prior to matching.

2.2.2 Assessing the accuracy or validity of a statistical match

29 Unlike in exact matching, the 'accuracy' of matched pairs is a concept not measured in statistical matching. There is no review or assessment of specific pairs for errors (Bernier and Nobrega 1999). Instead, practitioners of statistical matching examine the 'validity' of matched data. Rassler (2002, p. 30) proposes a framework to distinguish the four levels of validity matched data may achieve. From the easiest to hardest, the four tests of the validity of a matched data can be stated as follows:

1. Have the marginal distributions been preserved?
2. Have the correlation structures been preserved?
3. Have the joint distributions been preserved?
4. Have the individual values been preserved?

Level 1: Preserving marginal distributions

30 The minimum requirement of statistical matching is the preservation of the marginal distributions of the Z variables of File B, in the matched file. This is the easiest validity level to test, as one needs only to check the distributions of Z in the donor file and compare those against the distributions in the matched file. This validity is self-evident with constrained matching, but in the presence of "high-dimensional data structures and complex survey designs" even this level of validity may be hard to achieve (Rassler 2002 p. 32). Thus, in many studies, approximating the marginal Z distributions, rather than replicating the actual Z distributions, may be sufficient.

Level 2: Preserving correlation structures

31 This level requires that the correlation structure and higher moments of the X, Y and Z variables are preserved in the matched data, with $\tilde{\text{cov}}(X,Y,Z)$ = cov(X,Y,Z). (This is a consequence of but not as strong as CIA holding). This may be required if the interest after the matching is to examine the association of variables as measured by their correlation structure.

32 This level, together with levels 3 and 4, is harder to test and requires more than looking at the donor file and the matched file. It requires simulation studies or a third complete data source (or auxiliary information).

Level 3: Preserving joint distributions

33 This level requires that the true joint distribution of all variables be reflected in the matched file (Rassler 2002, p 30). Preserving the true joint distribution is important if subsequent analyses using the matched data require the use of all the X, Y and Z variables simultaneously and the analyses are sensitive to the joint distribution. Rassler pointed out that this level is met if the CIA holds.

Level 4: Preserving individual values

34 This refers to the reconstruction of the individual values, and therefore is the most difficult level of validity to achieve since we do not know the true values. The reproduction of the exact values may happen only if the common variables X determine in an exact manner the variable Z, as in identities or other functional dependence such as $Z = \alpha + \beta X_i$ or $Z_i = X_i^2$ (Rassler 2002 p. 30). Seldom is this level the main objective of statistical matching.

2.2.3 Constrained or unconstrained?

35 A fundamental question to ask is which method is more appropriate for a given objectived: unconstrained matching or constrained matching. Unconstrained matching is relatively simple and computationally easy, but its failure to maintain the marginal distributions of Z "can have a deleterious effect on the validity of the results of analyzing the matched file" (Cohen 1991 p. 65). Constrained matching is computationally demanding for large files, but in an era of increased computing power, this concern is becoming less important. Constrained matching's ability to replicate the marginal distribution of the Z variables, as required in the application discussed in section 3, makes it a more attractive method.

36 Rodgers (1984) reports that in Barr, Stewart and Turner (1982) and in Rodgers and deVol (1982), unconstrained matching can lead to "substantial distortions into the univariate and joint distributions of the Z variables" and that "regression analyses involving all three sets of variables indicated that unconstrained matches introduce more error than do constrained matches" (Rodgers 1984 p.99).

2.2.4 Choosing the matching variables

37 If preserving joint distributions is an important consideration given the modelling and analyses to be undertaken after the matching process, the matching variables to be chosen must be highly correlated with both the Y and Z variables. Cohen (1991 p. 67) suggests that to determine which X variables best predict Y and Z, a canonical correlation between Y and X can be done, and the X variable with the highest weights in the canonical correlation is chosen as the matching variables.

2.2.5 Defining a distance function and assigning weights to the Xs

38 Distance functions can be defined in several ways (Rassler 2002, p. 56).

39 One approach is defining it so that the absolute distance between any pair is minimised, using the 'city-block metric' formulation

$$d_{ij}^{CB} = \sum_{k=1}^{p} \left| x_{ik}^{A} - x_{jk}^{B} \right| . \qquad (6)$$

40 Where p is the number of x or common variables between files A and B, i is the unit observed in file A and j is the unit observed in file B.

41 The Euclidean distance function can also be used, defined as

$$d_{ij}^{E} = \sqrt{\left(x_i^A - x_j^B\right)' \left(x_i^A - x_j^B\right)}\,. \qquad (7)$$

42 In the practical application in section 3, the distance function used is of the Mahalanobis form, defined as

$$d_{ij}^{M} = \sqrt{\left(x_i^A - x_j^B\right)' S_x^{-1} \left(x_i^A - x_j^B\right)} \qquad (8)$$

where S_x is the estimated covariance matrix for the X variables.

43 In addition, the Mahalanobis distance function in section 3.6 is calculated using some user-defined weights corresponsing to the relative importance given to each matching variable.

3 Application of the conventional methods: creating a base file for a microsimulation model of the Pharmaceutical Benefits Scheme (PBS)

3.1 An overview of the PBS microsimulation model

44 The Commonwealth Government's Pharmaceutical Benefits Scheme (PBS) aims to provide Australians with timely, reliable and affordable access to necessary and cost-effective prescription medicines.

45 Patients are required to make a contribution to the cost of prescribed medicines listed on the PBS. Individuals and families eligible for certain federal government (Centrelink) pensions and allowances are able to access PBS medicines at concessional rates. The PBS also has 'safety net' arrangements to protect individuals and families from large overall expenses for PBS-listed medicines. The levels of patient copayments and the PBS safety net arrangements are referred to as the PBS policy settings. Patient copayments and safety net thresholds (SNTs) are revised annually in line with the consumer price index (CPI) from 1 January each year.

46 The majority of prescribed drug sales are covered by the scheme and, on average, the government subsidises patients to the extent of 84 per cent of PBS drug costs. Currently nearly 80 per cent of total government subsidies through the PBS accrue to concessional

patients — that is, those with the specified Centrelink cards[1] — and 20 per cent to general patients.

47 Finding ways of curbing government expenditure on the PBS while maintaining social equity and access to 'essential' medicines is at the centre of ongoing public debate. Since the early 1990s government expenditure on the PBS has grown at more than 10 per cent a year – well above the growth in the health budget (6 per cent) or the economy (4 per cent in terms of gross domestic product) while PBS copayments and safety net thresholds (that affect the cost to consumers) in general have increased only in line with inflation.

48 NATSEM models the Australian PBS using the microsimulation model MediSim. This model has two components: a Medicine Module that projects the total number of scripts and the average cost per script for 19 drug classes, and a Patient Module, whose main input dataset is at the person level (that is, each record is for an individual with links between family members).

49 MediSim simulates the current and future use and costs of PBS medicines under existing and different policy settings. It also estimates the distributional effects of policy changes. By altering the drugs included in the model, their assigned prices and script volumes, MediSim is capable, for example, of simulating the impact of inclusion of new drugs on to the list; restriction on the drugs listed on the scheme or on the pricing of drugs; increased restrictions on drugs by indication; increased use of generics at more competitive prices; or an increased emphasis on the quality use of medicines as reflected in changes in doctor prescribing behaviour; as well as changes to copayment and safety net arrangements.

50 The model could be used to provide answers to relatively simple issues such as the impact of expected changes in PBS subsidised drug prices and scripts over the next 5 to 10 years on government PBS outlays, or patient out-of-pocket expenditures and related revenues to industry. It could be used to measure the likely impact of, for example, the introduction of new PBS listed drugs, the effects of demographic and socio-economic changes upon outlays, or the distributional and revenue impact of certain changes in the rules of the PBS (eg. introduction of differential copayment levels).

51 Over the past year, NATSEM has been working to extend the modelling to include health outcomes. That is, both the *costs* and the *benefits* of pharmaceutical use will be modelled. The enhanced model could be used to help answer the more complex and difficult questions. For example, given an expected flow of new medicines on to the PBS

---

[1] These are the Pensioner Concession Card, the Commonwealth Seniors Health Card and the Health Care Card. For details, see the relevant Department of Family and Community Services fact sheets.

over the next decade or so, what will be the cost, how will this cost be shared between patients and government, and what will be the health benefits that accrue; or how would pharmaceutical usage and expenditure change in response to the earlier onset of diseases expected from the significant increases in obesity over the past five years amongst Australia's children and young adults? The significance of this type of modelling is that the proposed enhanced PBS model – if successfully implemented - will provide a more comprehensive picture of the contribution of pharmaceuticals to the Australian economy. It will also advance the debate on the sustainability of the Pharmaceutical Benefits Scheme by moving the discussion beyond the current focus on containing its cost. To realise this aim of including health outcomes in the model, the first step is to add diseases and health conditions, to the model base file.

3.2 Data issues for the MediSim basefile

52 NATSEM models the Australian PBS using the microsimulation model MediSim. This model is currently based on NATSEM's HES-based STINMOD 01A. The introduction of diseases and health conditions into the current model's dataset would require that we shift the base data of our model from the HES-based STINMOD, to the 2001 NHS (National Health Survey) CURF (Confidentialised Unit Record File) provided by the ABS.

53 The 2001 NHS provides details on the health of the Australian population, but has a number of limitations when applied to microsimulation modelling, particularly when applied to the PBS. The major limitation of the NHS in regard to modelling the PBS is the absence of family structure. While the NHS provides information at the person level and only limited details regarding family composition and inter-relations, the PBS must be modelled at the family level in order to effectively model the PBS safety net.

54 By statistically matching the NHS to the Household Expenditure Survey (HES) CURF we can use the rich health detail in the NHS and gain a family structure with information on every family member that is essential to modelling the safety net. Since the records to be matched involve sample surveys (rather than administrative data), and considering the incomplete coverage of families in the 2001 NHS, the matching of records will involve finding the closest statistical match between person records based on key variables, rather than exact matching of data records of the same persons.

55 Thus, NATSEM in cooperation with the ABS Methodology Division, embarked on a project to explore and develop the technique of statistical matching. The following sections investigate the statistical matching of the HES-based data called STINMOD 01A to the 2001 NHS.

56 As indicated above, NATSEM's main purpose for undertaking statistical matching is to synthetically create completely enumerated families, and making sure that persons on the NHS are sensibly assigned to families using the HES family structure. It is less of a concern for NATSEM to be able to use much, if any, of the Y variables in the resulting matched file.

With MediSim exclusively using the NHS (X, Z) variables, the concerns raised in section 2 regarding the potential issues with jointly using Y and Z variables is less of a concern.

3.3 Data to be matched

57 The matching undertaken here is between NATSEM's HES-based STINMOD 01A (which in this paper we interchangebly call this simply as "HES") and the 2001 NHS. STINMOD 01A contains around 18,000 person records and contains detailed income and expenditure information. Most of these variables are unique to STINMOD and using statistical matching terminology are the Y variables. STINMOD also contains more general information that will help in the statistical matching to the NHS, and these variables will be referred to as the X variables.

58 Each household that has been selected in the HES has a unique identifier for each household, family, income unit and persons. This hierarchy allows identification of persons to their correct income unit, family or household. As the PBS must be modelled at the family level, borrowing the family structure from the HES allows the proper modelling of the PBS' safety net. At this stage no variables from the HES will be used. The HES will purely be used for its structure.

59 The NHS contains around 27,000 person records. The NHS uniquely identifies persons and households. The persons in each household are randomly sub-sampled as follows:

- One adult (18 years or over)
- All children aged 0-6 years
- One child aged 7-17 years.

60 There is no information that allows for the unique identification of families or income units.

61 The NHS records have detailed information relating to the health of each person. This information is unique to the NHS and make up the "Z" variable list. Variables included in

the Z list that will be used initially in MediSim include: long-term conditions; medication usage; health insurance usage and health professional consultations[2]. The NHS also has some limited general information about each person. These "X" variables will be used extensively in the MediSim model and also to match the NHS to the HES.

62 To link records in both data sets we need variables that are common to both data sets and strongly related to the modelling area, in this case health. The list of X variables in the MediSim model is not large, and the strength of any relationship to health is likely to be more moderate than strong. The common characteristics include:

1. Age (6 groups)
2. Sex (2 groups)
3. PBS expenditure[3] (4 groups)
4. Income Unit type (4 groups)
5. Card holder status (2 groups)
6. Labour Force Status (4 groups).
7. Number of usual residents (6 categories)
8. Equivalised income decile (10 categories)

63 Appendix 1 shows the listing of the X, Y and Z variables.

64 The X variables from the NHS and the HES needed some modifications to ensure that their respective categories were equivalent. Matching records with the same or similar responses to the X variables form the basis for all methods of statistical matching in this paper.

---

[2] When the "health outcomes" module is added to Medisim this list is likely to grow.

[3] The NHS doesn't include this variable so the Self Assessed Health Status variable has been used as a proxy. This is discussed further in section 3.4.2.

3.4 Data modifications required for statistical matching

3.4.1 NHS data modifications

65 NHS' equivalent income decile variable includes a "Not Stated" category. This response is not compatible with the HES. This has meant that imputation was required for this variable for 4846 records (18 per cent) from the NHS data file. The imputation of these records was based on a polytomous logistic regression model that predicted the income deciles of the "Not Stated" category.

66 The concessional cardholder status of a person is important for calculating the cost of scripts. The NHS provides card status for that part of the population aged 15 plus. If a child belongs to a household that includes at least one adult with a card then the child will inherit that person's card status.

3.4.2 HES modifications

67 The original HES file required some alteration to turn it into a person-based file. This was achieved by running code that created a "kids" file and a separate "adults" file. The data sets were concatenated to produce the base HES file for matching.

68 As the HES has no variable on cardholder status, this was imputed using STINMOD information on receipt of FACS pension or allowance, plus other information such as age, gender, income level and family type. For each type of card, a pool of possible cardholders were identified based on receipt of particular pensions/allowances and eligibility for that type of card. (The only exception was for low-income health care card claimants, who qualified for cards solely based on their income.) Persons were randomly selected from the pool of possible cardholders in order to get the number of concession cardholders by type (HCC, CSHC, Pensioners) as close to FACS numbers as possible.

69 Equivalent income deciles were created using the OECD method for equivalising income. Some other minor modifications included: the income unit variable was altered to reflect the NHS version; the number of usual residents variable was capped at 6 plus as was the case on the NHS CURF; and if any HES householder was a cardholder then all other members of the household were also assigned a card.

70 The HES has precious little information relating to health from which we can match records to the NHS. There is household information on drug expenditure. Section 3.3 has PBS expenditure as a matching variable. The NHS does not have a corresponding variable, so a proxy was required. The NHS' self assessed health status (SAHS) was considered a

reasonable proxy. As the responses for these variables do not take the same distribution, the PBS expenditure variable has been allocated categories that match that of SAHS. Table 5 provides the category allocation of the PBS expenditure variable[4].

Table 5 PBS expenditure category allocation

PBS expenditure Values	(SAHS) Categories	SAHS category proportions	Cumulative %
1 (Least Expenditure)	Excellent	19.2 %	19.2 %
2	Very Good	33.3 %	52.5 %
3	Good	30.4 %	82.9 %
4	Fair	12.8 %	95.7 %
5 (Most Expenditure)	Poor	4.25 %	100 %

71 How reliable might proxying SAHS for PBS expenditure be? The 1995 NHS has information on both PBS usage and SAHS. The correlation between the usage of PBS drugs and SAHS was 0.41. Using the Spearman Rank Correlation statistic this result reduces to 0.3. As matching of the two variables will be by age, sex, cardholder status and income decile quintile a regression analysis was used to check on the joint significance of these variables when combined with SAHS to explain PBS usage. The adjusted R-square suggesting that self assessed health status and the other demographic variables explain around 41 per cent of variation in prescribed drug usage. It was anticipated that some of the unexplained variation might be accounted for by the NHS' drug usage window being only two weeks. Analysis was performed on a data set that removed the drug usage of people without serious conditions[5]. It was considered that this removed group were more likely to be using PBS drugs only intermittently. The correlation and regression results of these analyses fared no better when compared to the initial analyses.

72 The analysis of the link between the SAHS and PBS usage shows only a weak association. There were a number of alternatives to using SAHS as a proxy for prescribed drug expenditure. The number of doctor visits over the past fortnight was considered.

---

[4] Categories 4 and 5 were collapsed together for statistical matching purposes.

[5] As defined by the AIHW.

Unfortunately, due to the small time window, the majority of the population recorded no visits to the doctor. With so many people reporting no visits this variable is unlikely to provide a reasonable match to prescribed drug expenditure. Another alternative to the SAHS variable is the period of time since one last saw a doctor. One might expect that the longer the time spent not having seen a doctor could be related to less expenditure on prescribed medicines. Unfortunately the 1995 NHS does not have this variable so there is no empirical evidence to check this assertion. An a priori expectation would be that the link would be relatively weak, given the highly random nature of doctor visits for the large proportion of the population. A final alternative would be to combine SAHS and the time since seeing a doctor using principal components analysis. This option may be tried in the future, however time constraints don't permit this option at present.

73 The result that we only have one health related variable on the HES to match with the NHS is a concern. The situation is made worse by the above analysis suggesting the match is an imperfect one. Modelling any health Z variable with any Y variable will be relying upon one imperfectly matching health variable and various more general variables to create sensible statistical matches. Interpretation of results of any such analysis would require caution.

3.5 Statistical matching methods used

74 A two-step approach is used to statistically match the HES and the NHS. First, person records are subdivided into homogeneous groups or cells based on certain common variables. Next, persons belonging to the same group are matched together using a distance function. The cell groups are formed to ensure a certain standard is always maintained for the statistical match, while the distance function is a mathematical equation that attempts to more closely match individuals from the two surveys who fall within the same cell group. When using a small number of cell groups, the accuracy of matches can be improved by a properly formulated distance function.

3.5.1 Homogeneous cell groups

75 The methods used in this paper to statistically match are similar in that individual records are randomly matched within homogeneous cell groups. The methods differ by the method employed to match the individual records.

76 These homogeneous cell groups are based on records that have the same responses for the X variables. If the data sets were broken into cell groups based on all the possible combinations of the X variables listed in section 3.3 there would be some 115,200 unique cell groups. For our data set this means more cell groups than observations. Either

categories need to be collapsed, or certain X variables ignored. If the same people were on both datasets one could manipulate these cell groups to a point where you could exactly match records. The realities of sampling lead us to assume that the two samples have no overlap and records are only matched in a statistical sense.

77 Analysis of the NHS and HES surveys reveals that categories need to be rationalised significantly. Usage of only age, sex, cardholder status, income unit type and PBS expenditure (384 cell groups) revealed many empty cell groups in both the NHS and the HES. It is not until income unit type is removed that it is possible to obtain populated cell groups for all combinations (96 cell groups). The choice is now between using only a very small number of cell groups or using a larger number but having to merge cell groups where either the NHS or the HES cell group is empty. Either way, there are going to be some cells where both NHS and HES cell groups will be well populated containing hundreds of people while other cells may be as small as one person.

3.5.2 The distance function

78 If a cell based approach were the only method used to match then using a larger number of cell groups can only improve the quality of the matches. The statistical matching in this paper doesn't just use a cell groups based approach. Records within cell groups are not just randomly matched, rather some kind of sorting is done first. This sorting is based on a measure of closeness between individual records. Again the X variables are the basis for this measure of closeness.

79 This measure of closeness of potential matches, or distance function, is what complicates the process of choosing the right number of cell groups[6]. The cell groups are based on a hierarchy. Imagine if this hierarchy placed age and sex as the most important and income unit type as the least important. If we find that some cells in either the NHS or the HES are empty then some cells have to be collapsed. Given the hierarchy, cell groups could be joined at different ends of the income unit spectrum before any concessions will be made with the age variable.

80 If we have a more conservative number of cell groups and therefore very little collapsing, we can let the distance function determine the relative importance of the X variables. The distance function can be easily formulated to weight the relative importance of each X variable. In addition, using fewer categories increases the sample within each cell group, increasing the usefulness of the distance function.

---

[6] In the next section distance functions will be discussed in more detail.

81 In summary, the cell groups are there to ensure a certain standard is always maintained for the statistical match. Using more cell group categories will inevitably require the joining of some of these cell groups. Due to the hierarchical nature of the cell group system this can quickly lead to poor quality matches. When using a small number of cell groups the accuracy of matches can be improved by a properly formulated distance function. Regardless of the variable composition of the cell groups and distance function there is the further complication of collinearity between these matching variables. If two or more variables are included in the distance function that are strongly related the weights attached to each of these variables lose their meaning. The result being that varying the weights in the distance function may no longer produce the desired effect. With collinear variables in the cell groups we will be attempting to match combinations that we know are very unlikely to exist together. The result will be that some cell groups will have few or no persons.

3.6 Results

82 The above section discussed the basic methodology employed to match records. The separate datasets are firstly broken into homogeneous cells and then individual records are matched within these cells. The methods used in this paper vary by the way in which records are matched within the homogeneous cells. The methods we have used take on one of two forms, constrained or unconstrained. The unconstrained matching method leaves the original marginal distributions of both the donor and base data set variables changed. While the constrained method ensures that the original and base data set variables maintain their marginal distributions.

3.6.1 Unconstrained matching

83 Our initial and simplest approach is to match each HES record to the closest matching NHS record with replacement. Under this approach it is possible for the same NHS record to be matched with multiple HES records. The unconstrained matching procedure uses a distance function to determine the NHS record to be matched with a HES record. The selected match will minimise the following distance function:

$$\mathrm{d}_{i,k} = \sqrt{\sum_j a_{x_j} (X_{NHS,i,j,k} - X_{HES,i,j,k})^2 / \sigma^2_{x_j,HES}} \qquad (9)$$

84 The subscript i relates to person records, j to the matching variable, and k to the cell group. $\sigma^2_{x_j,hes}$ is the variance of the jth matching variable. a_{x_j} is the user defined relative importance, or weight given to each matching variable. The matching variables that were

used in the distance function were; age, number of usual residents and equivalised income deciles. The age variable in the distance function has 16 possible categories, somewhat more than the 6 categories employed in the cell groups.

85 In the event of multiple records minimising the distance function the first record is taken. As records are ordered by their person identification number this is equivalent to a random selection.

86 Equation (9) has provided a means for ensuring that records from the basefile HES are linked to records in the NHS that are statistically similar. As this method is "with replacement" once a NHS record has been selected for matching with a HES record it is not precluded from being matched with other HES records.

87 Selecting NHS records for matching in such a manner ensures that the "match quality" is high. When dealing with sample surveys with relatively small samples it can be difficult to match many of the observations, but allowing the NHS records to be replaced after they have been matched to some other HES record significantly reduces this problem.

88 Table 6 shows the accuracy of the match with respect to the X variable age. Age was a cell group variable and this ensured a certain level of accuracy[7]. In the table below age values from the NHS between 40 and 64 will always be linked to HES records within that same age band. The distance function tightens the match very successfully for the unconstrained method. Consider the HES age group 60 to 64 in the "merged" data. 95 per cent of HES records have been matched to NHS records with the correct age group.

[7] The results are based on unconstrained matching where the distance function used weights of 0.5 for age, 0.25 for equivalised income decile and 0.25 for the number of usual residents in the household.

Table 6 Unconstrained matching age allocation

	NHS age grouping															
	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 plus
Merged (HES age grouping)																
0 - 4	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5 - 9	.	90	10	.	.	.	.	.	.	.	.	.	.	.	.	.
10 - 14	.	15.7	84.3	.	.	.	.	.	.	.	.	.	.	.	.	.
15 - 19	.	.	.	88	12	.	.	.	.	.	.	.	.	.	.	.
20 - 24	.	.	.	37.6	62.4	.	.	.	.	.	.	.	.	.	.	.
25 - 29	.	.	.	.	.	78.8	17.6	3.6	.	.	.	.	.	.	.	.
30 - 34	.	.	.	.	.	11.8	74.7	13.5	.	.	.	.	.	.	.	.
35 - 39	.	.	.	.	.	3	12.4	84.6	.	.	.	.	.	.	.	.
40 - 44	.	.	.	.	.	.	.	.	82.5	13.4	3.6	0.4	.	.	.	.
45 - 49	.	.	.	.	.	.	.	.	15.2	72.2	11.2	1.4	0.1	.	.	.
50 - 54	.	.	.	.	.	.	.	.	2.7	12.3	75	8.8	1.2	.	.	.
55 - 59	.	.	.	.	.	.	.	.	0.5	3.1	13	72.3	11.1	.	.	.
60 - 64	.	.	.	.	.	.	.	.	0.3	0.6	3.2	12.2	83.7	.	.	.
65 - 69	.	.	.	.	.	.	.	.	.	.	.	.	.	89.2	8.7	2.1
70 - 74	.	.	.	.	.	.	.	.	.	.	.	.	.	7.4	85.7	6.9
75 plus	.	.	.	.	.	.	.	.	.	.	.	.	.	4.9	5.8	89.3

89 Section 2.2 discussed the various validity levels that are required for statistical matching to provide reliable outcomes. As a minimum standard it should be expected that marginal distributions should remain intact. If this standard was to be met the means and standard deviations of the Z variables in the NHS should remain unchanged by the matching process. Unconstrained matching does not satisfy this condition. This condition is broken for two reasons: the weights attached to the Z variables are the weights that belong to the HES; and it is often the case that certain NHS observations are either never selected or are selected multiple times. Table 7 shows the selected NHS variable's marginal distributions are relatively close to the original NHS'.

Table 7 Marginal distributions: NHS versus matched data using unconstrained matching

	Index of Relative Socioeconomic Disadvantage (%)					
	1	2	3	4	5	total
NHS	17.3	19.4	19.2	22.8	21.2	100
Matched data	16.9	18.9	19.3	22.7	22.1	100

	Number of times consulted GP

	0	1	2	3	total
NHS	78.2	17.6	3.4	0.8	100
Matched data	78.6	17.6	3	0.8	100

90 The problems with unconstrained matching become more apparent when working with data at a finer level of disaggregation. The MediSim model will rely heavily on the use of the conditions data in the NHS. The NHS shows whether or not individuals have any of 94 different long-term conditions. Analysis was done on each of these conditions and a comparison was made between total number of cases for the unconstrained matching-based matched file and the original NHS data. The ratio of the total numbers for each condition for the matched data and the original data was computed. A value of 1 implying that the unconstrained matched file provided a perfect representation for a given condition. For all 94 conditions, the unconstrained matched file averaged 0.96, but ranged from 0.24 to 1.32. Such a result is not considered adequate and obviously alternative methods needed to be found.

3.6.2 Constrained matching

91 Ideally we need a method that can provide good quality matches on the X variables, but ensure that marginal distributions are maintained. This may be achieved by constrained statistical matching, the objectives and conditions of which were laid out in section 2.1.2. As outlined in section 2.1.2, for constrained matching, linear programming (LP) can be used to match records in each data set in such a way that the distance function in equation (9) is minimised, subject to the constraint that the weights attached to the matched file preserve the weights of each of the separate files. Matching in such a way has the special property that the marginal distributions of both the HES and the NHS will be preserved.

Table 8 Linear programming tableau

HES/NHS	25		20		20		35	
30	w_{ij}	*1.2*	w_{ij}	*2.1*	w_{ij}	*4.3*	w_{ij}	*0.8*
20	w_{ij}	2.2	w_{ij}	1.1	w_{ij}	6.3	w_{ij}	*0.9*
40	w_{ij}	*3.3*	w_{ij}	*0*	w_{ij}	*0.9*	w_{ij}	*2.4*
10	w_{ij}	*1.4*	w_{ij}	*2.6*	w_{ij}	*4.3*	w_{ij}	*1.8*

92 Table 8 describes the matching process undertaken when using the linear programming approach. In this illustrative case we are looking to match four HES observations with a total weight of 100 to four NHS observations with a total weight also of 100. The first column represents the original weights for each HES record. The first row represents the original NHS weights. The numbers in italics represent the distance function value between each HES and NHS record. The linear programming problem is to assign the weight of each HES record amongst the four NHS records. The weight from each HES record that is assigned to each NHS record is the weight that applies to the new matched file, w_{ij}. This assignment should be applied so there is no "slack" in the allocation. This implies that the total supply of 100 is fully exhausted by the demand of 100 from the NHS records. The linear programming solution ensures that the resulting assignment minimises the distance function. This application of linear programming is also known as the "transportation problem".

93 The linear programming approach requires the sum of HES and NHS weights to be equal. This ensures a "balanced" problem. If a weight is interpreted as the number of people a record represents then an unbalanced problem leads to either people in the HES or the NHS not being matched. This situation was described as "slack" earlier. As this method is applied to each of the homogeneous cells, adjusting the weights of either the HES or the NHS to ensure a balanced solution will change the relative importance of each cell. The weights in this particular application were always re-weighted to the NHS population. This will mean that marginal distributions will not hold for the HES variables, the marginal distributions will not change for the NHS variables. Table 9 provides the extent to which selected variables in the matched file equate with the original HES data.

94 In the unconstrained matching section the marginal distributions for the 94 long-term conditions in the NHS were discussed. The statistically matched file often over or under-reported the incidence of these conditions. The constrained matching method ensures that the incidence in the matched file is identical to that of the original NHS file[8].

[8] As the linear programming method uses integer programming the weights attached to the matched file were rounded. This has led to some very small discrepancies in marginal distributions.

Table 9 Marginal distributions of selected HES variables

	Original HES	*LP Merged data*
Average tax ($/ft)	608	583
Average Dwelling Value	151172	148204
Average Total Income ($/ft)	1074	1030
PBS expenditure ($/ft)	2.82	3.15
Card status	*Yes*	*No*
HES original	42.9	57.1
LP Merged	34.2	65.8

95 Table 6 illustrated that when matching with replacement (i.e. using unconstrained matching) the quality of the match is high with respect to the matching variables. The linear programming method only selects records from the NHS without replacement[9]. This is expected to reduce the quality of matches. Tables 10 to 12 provide a measure of the "closeness" of the match between the variables in the distance function. The weights that have been attached to the distance function are unchanged from those used to produce the results in Table 6. Table 10 indicates that the closeness of the age match is not as robust as that of the WR matching. The results are still promising with very few records in the HES being matched to NHS records where the age categories are more than 1 group apart. Tables 11 and 12 show the results for the closeness of the matches for income deciles and the number of usual residents. The income results are quite poor with many HES records being matched with NHS records more than 2 categories apart. The number of usual residents match shows a relatively close match.

[9] For the LP approach each record was "exploded" so that a record was repeated to the extent of its weight. The selection without replacement refers to this "exploded" data set. It is quite possible that repeats of the same record will be matched multiple times.

Table 10 Quality of age match for LP matching

	NHS age grouping															
	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 plus
Merged (HES age grouping)																
0 - 4	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5 - 9	.	87.3	12.7	.	.	.	.	.	.	.	.	.	.	.	.	.
10 - 14	.	10.9	89.1	.	.	.	.	.	.	.	.	.	.	.	.	.
15 - 19	.	.	.	87.9	12.1	.	.	.	.	.	.	.	.	.	.	.
20 - 24	.	.	.	24.8	75.2	.	.	.	.	.	.	.	.	.	.	.
25 - 29	.	.	.	.	.	77.2	19.5	3.3	.	.	.	.	.	.	.	.
30 - 34	.	.	.	.	.	13.5	73	13.5	.	.	.	.	.	.	.	.
35 - 39	.	.	.	.	.	1.4	17.9	80.7	.	.	.	.	.	.	.	.
40 - 44	.	.	.	.	.	.	.	.	70.1	19.3	7.8	2.8	.	.	.	.
45 - 49	.	.	.	.	.	.	.	.	12.4	58.3	21.8	6.5	1.1	.	.	.
50 - 54	.	.	.	.	.	.	.	.	4.1	11.7	56.6	22.1	5.6	.	.	.
55 - 59	.	.	.	.	.	.	.	.	1.1	3	10.4	59	26.5	.	.	.
60 - 64	.	.	.	.	.	.	.	.	0.3	0.9	2.2	14.7	81.8	.	.	.
65 - 69	.	.	.	.	.	.	.	.	.	.	.	.	.	69.4	22.7	8
70 - 74	.	.	.	.	.	.	.	.	.	.	.	.	.	9	62	29
75 plus	.	.	.	.	.	.	.	.	.	.	.	.	.	2.6	8.1	89.3
All	9.3	9	7.3	6.8	4.9	6.4	8.4	8	7.3	6.7	6.3	5	4.3	2.7	3	4.5

Table 11 Quality of income decile match for LP matching

	HES									
NHS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	34.9	9.6	7.1	12	14.2	12.3	2.9	2.2	1.4	3.4
2	11.8	30.5	17.1	15.5	9.2	7.1	2.5	1.3	2.2	2.8
3	5.3	10.7	27.8	15	16.3	10.8	2.3	2	3.4	6.4
4	3	6.5	2.1	22.9	15.4	11.9	8.9	6	7.9	15.3
5	3.5	5.8	3.2	4.3	23.3	13	11.4	11.8	10.5	13.2
6	2.8	6.4	3.6	4.1	6.5	30.8	15.7	12.1	8.8	9
7	2.3	6.5	3.9	6.2	5.1	7.9	42.2	13.2	6.3	6.3
8	1.8	4.6	3.6	5.3	4.3	4.5	8.8	46.8	13	7.2
9	1.2	3.2	4.6	8.2	5.9	4.9	7.5	7	45.5	11.9
10	0.6	1.4	3.8	7.9	6.5	2.9	5.2	5.4	12.3	54

Table 12 Quality of number of usual residents match for LP matching

	NPER - NHS					
NPER - HES	1	2	3	4	5	6
1	81.3	14	3.2	0.8	0.4	0.2
2	4.4	81.8	10.2	2.9	0.5	0.1
3	2.3	17.5	60.9	17.7	1.1	0.4
4	0.6	7.3	11.5	72.4	7.4	0.7
5	0.9	1	1.8	16	66.5	13.8
6	2.4	1.1	1.6	3.2	19.6	72.1
All	11.8	26.2	17.2	25	13.3	6.5

96 To improve the match of any of the three X variables that make up the distance function the weights in the distance function need to be altered. If the statistical match requires income to be well matched then the relative importance of this variable should be increased. Naturally, any such changes are likely to impact adversely upon the closeness of match of other X variables in the distance function.

97 The appendices show the results of altering the distance function weights to the constrained matching method. In Appendix 2 the closeness of match results have been displayed for income decile. Surprisingly the results hardly differ regardless of whether the weight attached to income is .67 or .33[10]. One possible explanation for this odd outcome is that age and income are strongly correlated. The weight mix of two correlated variables may have little impact on the closeness of the match for either variable.

98 Appendix 3 again uses the income decile variable. This time a comparison has been made between a distance function where the weight attached to income is one and all other variables zero, the second, where age has a weight of one and the other variables zero. This provides us with the boundaries of matching closeness that can be obtained for the income variable. The results for income are not promising, even for the extreme case where the distance function only uses the income decile variable we still don't achieve particularly close matches for income deciles.

99 Appendix 4 shows these same results as Appendix 3 except considering the age variable. Again, the zero weight shows very poor results but this time when weight is set to one the match for age is very close.

100 For some applications of statistical matching it is quite likely that the researcher will require that the statistically matched file be well matched to the income decile variable. In this application income was not well matched using the constrained matching. The solution to this problem is to use some form of the income decile variable as a cell

[10] The weight attached to the number of usual residents in this simulation was 0.

group variable. Keeping in mind the problems discussed in section 3.4 the inclusion of an income variable will probably need to be at the expense of some other cell group variable, either in terms of collapsing categories or the removal of a variable.

3.6.3 Creation of synthetic families

101 As discussed in Section 3.3, the NHS is limited in household structure details. Using the household structure underlying the HES in the matched HES-NHS file overcomes these problems. The statistical matching process synthetically creates families. This section investigates the efficacy of such a process.

102 The HES records are made up of complete and real families, as each family is fully enumerated. As the matching process is at the individual level, the NHS records that are matched to these HES families are most likely to be drawn from different families. It would make most sense to be matching at a more aggregated level, such as the family or household level as these units are likely to be more homogeneous in their characteristics. Section 3.1 outlined the various reasons for why this would not be possible.

103 One of the variables used to create the homogeneous cell groups is the income unit type. The income unit type is defined in both the HES and the NHS as either: single person; single person with children; couple without children; couple with children, not known[11]. In the creation of cell groups, singles and not known are grouped together and the rest are placed into a second group. This does mean that the NHS individuals that are matched to a given HES family are not only from different families but potentially different family types. Table 13 describes the matches that are achieved using constrained matching[12].

[11] The income unit type is used in preference to the family type variable as the former better describes the family groupings used for social security purposes.

[12] Due to rounding some rows do not add to 100 per cent.

Table 13 HES – NHS Income unit type matches

NHS matched income unit type	HES Income unit type (%)					
	Single	Single + Dependents	Couple only	Couple + dependents	Not known	Total
Single	81				19	100
Single + Dependents		45	15	40		100
Couple only		3	81	16		100
Couple + dependents		6	7	87		100
Not known	56			4	41	100

104 The table does show that the income unit type allocation is far from perfect. Had income unit type categories not been collapsed the above table would have shown the correct allocation.

105 Should we improve the income unit type allocation? For reasons given in section 3.5.1, increasing the detail of the income unit type variable will be at the expense of other matching variables. At this point it must be clear why we are matching. The purpose of matching in this case study is to create a file structure amenable to modelling family PBS expenditure. Family PBS expenditure depends on the health and card status of individuals in a family. The variables that have been used in the cell groups and the distance function attempt to account for these factors. Without any formal empirical analysis it is difficult to determine what factors really do most strongly determine family PBS expenditure. As a result the current mix and usage of matching variables is open to debate and could do with further research.

4 Conclusion

106 The statistical matching of separate data sets is traditionally performed to add important information that doesn't exist on any of the separate data sets alone. In this case study the HES data set has been merged onto the NHS. This has not only added

information to the NHS but has overcome some of the shortcomings that the NHS poses to microsimulation modelling by creating complete synthetic families.

107 Two possible matching procedures have been compared, unconstrained matching, where NHS records can be matched to HES records with replacement, and constrained matching where linear programming was used to ensure that marginal distributions of at least the NHS remained constant. Both procedures have the ability to match relatively closely on the linking X variables. Only the constrained matching can guarantee that marginal distributions will remain unchanged.

108 Statistical matching is as much art as science. There is no simple recipe for creating a statistically matched file that will satisfy all requirements. For the methods used here there are practical trade-offs that need to be considered.

109 The first trade-off is the relative usage of cell groups and the distance function. Having a great number of cell groups does provide control to which records are matched. When merging of cell groups is required to overcome empty cell groups, the hierarchical nature of cell groups may lead to poor matches. While the distance function doesn't have the hierarchy problem, one loses a degree of control over the process. The constrained matching example illustrated this trade-off well. Income decile was only included in the distance function. The match quality in terms of income was quite poor regardless of the weight given to income in the distance function. The key point here is that if a variable is thought to be of great importance to the matching of data sets then it should be included at least as a cell group variable, and if possible at a finer level of detail in the distance function.

110 The second trade-off is the relative importance of the matching variables, both in the distance function and the cell groups.

111 Making decisions on these important trade-offs will typically come down to the individual researchers own judgement and be done on a case-by-case basis. A further complicating matter is collinearity between matching variables. If two or more variables are included in the distance function that are strongly related the meaning attached to the distance function weights may be lost. Using multicollinear variables in the cell groups is likely to lead to some cell groups with few or no persons.

112 Much caution needs to be used in interpreting any results of a statistically matched file where Y and Z variables are being used in the same model. If they are, there should be sufficient evidence indicating that both joint probability distributions and correlation structures of the X,Y,Z variables are preserved. Unfortunately, testing for these is relatively difficult to do, and is beyond the scope of this paper.

113 NATSEM intends to use the matched file that was estimated using constrained matching. Essentially, the person records in the original NHS were reshuffled into different families based on the HES family structure, such that information on every family member (that is essential to modelling the safety net) is available. Given that the original NHS persons records have now been reconstructed into complete families, NATSEM's intention is to use only the variables from the NHS and not to use the Y, Z relationships. In essence, this means that {X,Z} distributions remain intact, and in fact, with the exception of family structure, individual values in the NHS are preserved in the statistically matched file.

5 Further research

114 In this paper, we focussed our attention on two approaches to statistical matching, the unconstrained and constrained methods. Other approaches are possible, many of which are considered under the heading of multiple imputation approach and use of auxiliary information (Rubin 1986, Paass 1986, Rassler 2002). The problem with stat matching is that it is not possible to attach a meaningful measure of accuracy to the match to show how much variation there would be across all possible matches under a particular matching scheme. Alternative matching approaches such as multiple imputation do have this advantage. These methods are supposedly better in taking full account of the information in the input files. Thus it may be possible to see whether the joint distributions of X,Y,Z variables, or their correlation structures, are preserved after the matching. There may be methods too that dispose of the conditional independence assumption, by using auxiliary information to avoid the CIA (e.g. Singh et al 1993). These alternative methods need to be considered especially if the X, Y and Z variables in a matched data set are going to be used simultaneously in any modelling exercise.

References

Barr, R.S. and Turner, J.S. , 1978. "A new, linear programming approach to microdata file merging", *1978 Compendium of Tax Research*, Office of the Treasury, Washington, D.C.

Barr, Stewart, W.H. and Turner, 1982. *An Empirical Evaluation of Statsitical Matching Strategies*, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

Cohen, M. L. (1991), "Statistical Matching and Microsimulation Models," in *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Vol. II. Technical Papers, eds. C. F. Citro and E. A. Hanushek,

Ingram, D. D., O'Hare, J., Scheuren, F., and Turek, J. (2000), "Statistical Matching: A New Validation Case Study," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 746-751.

Kadane, J., 2001, Some Statistical Problems in Merging Data Files, *Journal of Official Statistics*, Vol. 17, No. 3, pp. 423-433.

Moriarity, C., and Scheuren, F., 2001. *Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure,* Journal of Official Statistics, Vol. 17, No. 3, pp.407-422.

Moriarty, C., and Scheuren, F., 2001. 'Statistical Matching: Pitfalls of Current Procedures', in *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.

Paass, G. (1986), "Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information," in *Microanalytic Simulation Models to Support Social and Fiscal Policy*, eds. G. H. Orcutt, J. Merz, and H. Quinke, Amsterdam: North-Holland, pp. 401-420.

Radner, D. B., Allen, R., Gonzalez, M. E., Jabine, T. B., and Muller, H. J., 1980. "Report on Exact and Statistical Matching Techniques," *Statistical Policy Working Paper 5*, Federal Committee on Statistical Methodology, Office of Management and Budget, U.S. Government Printing Office.

Rassler, Susanne, 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches,* Springer-Verlag.

Rodgers, W., 1984. An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics,* Vol. 2, No. 1, January 1984, pp. 91-102.

Rubin, D., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, Vol. 4:86-94.

Scheuren, F., and Winkler, W. E. (1993), "Regression Analysis of Data Files That are Computer Matched," *Survey Methodology*, 19, 39-58.

Sinclair, M., Potter, F., and Carlson, B.L., 2001. 'Statistical Matching Techniques for a Household Health Insurance Survey', in *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.

Singh, A.C., Mantel, H.J., Kinack, M.D., and Rowe, G., 1993. 'Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption', *Survey Methodology*, Vol. 19, No. 1, June 1993.

Appendix 1. Listing of X, Y and Z

Y (HES)	X (In both HES and NHS)	Z (NHS)
	Age	
	Sex	
	Expenditure on prescribed medicine[a]	
	Income unit type	
	Concession card[b]	
	Labour force status	
	Equiv. income decile	
	No. of usual residents	
	SEIFA	
Weekly total income of the income unit, equal to the sum of private and transfer income.		Self-assessed health (also used as proxy for expenditure on prescribed medicine)
Weekly disposable income of the income unit		Long-term conditions and national health priority conditions
Weight attached to reference person. Also used as the weight for the family		Actions taken (including medication)
Weight attached to each person		Risk behaviour
Number of children aged between 0 and 4 in the income unit[c]		Immunisation
Number of children aged between 5 and 9 in the income unit[c]		Country of birth
Number of children aged between 10 and 12 in the income unit[c]		Main language spoken at home
Number of children aged between 13 and 14 in the income unit[c]		Highest educational qualification
Number of children aged 15 in the income unit		Labour force status
Type of FACS pension or allowance received		SEIFA
Expenditure on prescribed medicine		Location
Expenditure on doctor consultations		Private health insurance
Expenditure on hospitalisation		Government concession card
		Days away from work
		Other days of reduced activity
		Person weight
		Number of times admitted to hospital
		Number of nights in hospital
		Number of times consulted GP
		Number of times consulted specialist

		Time since last consulted a doctor Number of times consulted dentist Time since last consulted dentist Body Mass Index (BMI)

[a] On the NHS, the proxy variable for this was self-assessed health status.

[b] On the HES, a proxy variable was imputed based on eligibility for government pensions and allowances.

[c] Used to create individual records on children.

Appendix 2

LP 5 cells - income 0.33 age 0.67

Equivalent Income Decile (%) Correlation = 0.60

NHS \ HES	1	2	3	4	5	6	7	8	9	10
1	36.2	3.3	10.7	15.5	20.4	8	2.7	1	1.8	0.4
2	14.2	32.9	19	14.5	9.7	8.1	1	0.3	.	0.3
3	5.7	9.5	33.7	13.6	14.3	13.8	8	1.1	.	0.2
4	5	5.5	2.2	27.6	13.8	13.7	17.7	14.1	0.4	0
5	4.6	2.9	2.2	3.4	31.2	12.1	15.2	18.1	10.3	.
6	0.3	5.5	4.8	4.7	5	38.9	13.6	13.2	8.2	5.9
7	0.8	0.2	8.5	6.6	5.8	2.9	50.4	11.3	9.4	4.1
8	1.1	0.9	0.2	12.2	6.3	1.6	5.6	58.8	8.8	4.6
9	1.5	1	0.7	0.7	16	2.4	5.3	6.1	62.5	4
10	0.5	2.5	2.4	2.9	2.1	7.1	7.5	7.7	13.8	53.5

LP 5 cells - income 0.67 age 0.33

Equivalent Income Decile (%) Correlation = 0.58

NHS \ HES	1	2	3	4	5	6	7	8	9	10
1	37.1	3.8	9.3	12.7	20.1	7	3.8	2.7	0.9	2.6
2	8.6	36.9	16.3	13.8	9.1	10.8	1.5	1.5	0.3	1.2
3	5.5	9.9	32.6	13.5	12.5	12	10.3	1.2	0.4	2
4	4.5	7.3	2.4	26.2	13.6	12.3	15	14.6	2.3	1.8
5	4.1	7.2	3.1	3.4	28.1	12.6	12.3	15.2	13.7	0.3
6	1.1	8.8	5.3	5.6	3.6	33.6	15.3	8.3	7.9	10.4
7	1.7	2.8	10.2	8.4	4	3.6	43.2	8.8	8.2	9.2
8	1.9	2.6	0.7	13.3	4.1	2.5	3.6	46.6	14.1	10.5
9	3.3	4.4	2.7	1.9	12.1	4.4	1.9	3.1	52	14.1
10	2	6.4	5.4	3.8	0.6	5.3	2.7	1.2	8.4	64.1

Appendix 3

LP 5 cells - age 1

Equivalent Income Decile (%) Correlation = 0.38

NHS \ HES	1	2	3	4	5	6	7	8	9	10
1	18.4	16.5	11.2	12.8	13.6	10.6	5.7	3.9	3.9	3.3
2	17.3	21.8	14.9	14.3	11.8	7.9	4.4	2.9	2.3	2.4
3	11.9	13	14.1	15.5	14.7	11.7	6.6	4.6	4.4	3.5
4	6.8	8.4	9.2	12.1	12.2	13	9.9	9.2	10.3	8.9
5	6.3	6.6	6.8	8.3	9.5	10.1	12.6	12.9	13.4	13.7
6	5.5	6.5	6.3	7.3	7.3	10.2	13	13.7	14.8	15.4
7	5.5	6.6	6.8	6.9	6.9	10.2	11.8	14.8	14.1	16.3
8	3.6	4.9	4.8	6.7	7.6	10.1	14.8	15.1	16	16.2
9	3.7	5.2	5.5	7	9	11	15.2	14.8	14.8	14
10	2.9	4.1	5	6.7	9.2	11	14.7	14.7	15.2	16.5

LP 5 cells - income 1

Equivalent Income Decile (%) Correlation = 0.65

NHS \ HES	1	2	3	4	5	6	7	8	9	10
1	41.6	5.2	13.9	30.2	0.3	1.4	2.1	5.3	0	0
2	9.9	38.2	13.1	10.4	25.4	0.2	0.4	0.5	1.9	.
3	8.9	11.4	26	7.6	15.5	29.4	0	0.3	0.3	0.6
4	5.6	7.5	1.8	18	14.2	19	33.4	0.1	0.1	0.4
5	.	13.7	3.1	2.8	24.1	10.5	16	29.7	0	.
6	0.4	0.8	16.4	5.4	2.5	26.8	11.6	14.1	22	.
7	2.2	2.1	0.2	18.4	3.4	2.6	34.5	7.6	10.1	19
8	2.5	2.8	0.4	0.1	16.9	2.7	2.7	41.7	11.6	18.5
9	.	10.7	2.9	2.2	0.6	10.9	2.7	1.7	53.7	14.7
10	0.3	0.2	11.2	5.4	1.6	0.2	4.7	4.4	8	63.9

Appendix 4

LP 5 cells - 1 age

Correlation = 0.99

Merged (HES age grouping) \ NHS age grouping	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 plus
0 - 4	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5 - 9	.	90.9	9.1	.	.	.	.	.	.	.	.	.	.	.	.	.
10 - 14	.	5.1	94.9	.	.	.	.	.	.	.	.	.	.	.	.	.
15 - 19	.	.	.	94.3	5.7	.	.	.	.	.	.	.	.	.	.	.
20 - 24	.	.	.	18.4	81.6	.	.	.	.	.	.	.	.	.	.	.
25 - 29	.	.	.	.	.	84.4	10.5	5.1	.	.	.	.	.	.	.	.
30 - 34	.	.	.	.	.	6.1	90	4	.	.	.	.	.	.	.	.
35 - 39	.	.	.	.	.	2.7	7.6	89.7	.	.	.	.	.	.	.	.
40 - 44	.	.	.	.	.	.	.	.	82.5	10.8	6.7	.	.	.	.	.
45 - 49	.	.	.	.	.	.	.	.	6.4	78.8	6.6	8.1	.	.	.	.
50 - 54	.	.	.	.	.	.	.	.	1.6	1.3	76.2	8.8	12.1	.	.	.
55 - 59	.	.	.	.	.	.	.	.	.	0.8	3.8	80.9	14.5	.	.	.
60 - 64	.	.	.	.	.	.	.	.	.	.	3.2	3.5	93.3	.	.	.
65 - 69	.	.	.	.	.	.	.	.	.	.	.	.	.	75	10.8	14.1
70 - 74	.	.	.	.	.	.	.	.	.	.	.	.	.	6.3	76.4	17.3
75 plus	.	.	.	.	.	.	.	.	.	.	.	.	.	3.9	4.7	91.4

LP 5 cells - 1 income

Correlation = 0.96

Merged (HES age grouping) \ NHS age grouping	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 plus
0 - 4	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5 - 9	.	56.3	43.7	.	.	.	.	.	.	.	.	.	.	.	.	.
10 - 14	.	54.3	45.7	.	.	.	.	.	.	.	.	.	.	.	.	.
15 - 19	.	.	.	64.4	35.6	.	.	.	.	.	.	.	.	.	.	.
20 - 24	.	.	.	50.9	49.1	.	.	.	.	.	.	.	.	.	.	.
25 - 29	.	.	.	.	.	32.3	33.2	34.5	.	.	.	.	.	.	.	.
30 - 34	.	.	.	.	.	29.7	33.6	36.7	.	.	.	.	.	.	.	.
35 - 39	.	.	.	.	.	28.1	34.8	37.1	.	.	.	.	.	.	.	.
40 - 44	.	.	.	.	.	.	.	.	27.4	24.2	20.5	15.8	12.2	.	.	.
45 - 49	.	.	.	.	.	.	.	.	26.6	23	20.8	15.2	14.5	.	.	.
50 - 54	.	.	.	.	.	.	.	.	26.2	22.4	21.3	15.2	15	.	.	.
55 - 59	.	.	.	.	.	.	.	.	25.6	21.1	19.3	17.5	16.5	.	.	.
60 - 64	.	.	.	.	.	.	.	.	25.5	17.4	17.6	16.9	22.7	.	.	.
65 - 69	.	.	.	.	.	.	.	.	.	.	.	.	.	29	27.6	43.3
70 - 74	.	.	.	.	.	.	.	.	.	.	.	.	.	29.2	29.4	41.3
75 plus	.	.	.	.	.	.	.	.	.	.	.	.	.	27.5	28.6	43.9