SHARE FINAL OUTCOMES [posted as supplied by authors]

Data acquisition, linkage and analysis protocol

Gillian Raab and Marion Henderson

1 Introduction

The analysis of SHARE follow-up data was planned at the time of our first application to fund the study in 1991. At that time we obtained written permission from the Information and Statistics Division of the NHS in Scotland (ISD) that we would be able to obtain NHS data on conceptions and terminations of the young women in SHARE and control schools. Exact details were not specified, but we understood that this would be done via probability matching to data that we would supply to ISD and that we would not have access to individual data.

As the time approached when our second cohort of young women would all reach their twentieth birthday we entered into negotiations with ISD as to what data we would be able to obtain. After discussions with ISD as to what would be possible, we agreed that they would carry out the record linkage and supply the data according to a written agreement that is attached here as Appendix 1. It was agreed that aggregate data would be made available, but that we could obtain data that would subdivide the pupils within a given school, provided there were few, if any, subgroups of fewer than 5 pupils. The aggregation we requested is detailed in a letter from Marion Henderson to ISD attached as Appendix 2.

This document explains the rationale for our selection of this aggregation and sets out the analysis protocol that was used to analyse the data.

2 Choice of aggregation classes

We had a very wide range of data available from the cohort of young people from the questionnaires that were completed as part of the SHARE study. These data have been extensively analysed in publications (see http://www.msoc-mrc.gla.ac.uk/share/findings/subject-MAIN.html). We wanted to obtain data from ISD disaggregated by factors that would be needed for

- to adjust the analysis of the outcomes for individual factors
- to allow us to report conceptions, births and terminations data on subgroups of the population that might be of interest.

For the first of these we wanted to select factors that would be predictive of the outcomes. It was also essential that we selected only data obtained pre-intervention. We used our work on the predictors of sexual behaviour to guide this choice. Since one of the strengths of using routine

data at follow-up was the completeness of the data, it was important that we select data that were completely, or almost completely recorded on our records. We selected two factors that fulfilled these criteria 1.Whether pupil left school at the earliest opportunity and 2. Social class of father (or mother if no data data available for the mother, coded as manual or non manual. When social class data were incomplete data was imputed from the home post-code. A cross-tabulation of these two factors provide too many small cells, particularly in the non-manual early school-leavers category. To overcome this we decided to ask for data subdivided by early leaver status, and then broken down by manual/non-manual social class for those who were not early leavers.

Additional factors of interest were the cohort within the study and the age at which the conceptions took place. To avoid small cells data were requested as three separate aggregations that are detailed in Appendix 2.

3 Analysis plan

The analysis plan was heavily constrained by the restrictions of the aggregated data. These identified our four outcome variables. The primary outcome variable (as in our original trial protocol) was the rate of terminations for share and control women. Some women would contribute more than one termination so we also obtained data to identify the number of young women with any termination during the follow-up period. The secondary outcome was defined similarly, but for conceptions rather than just abortions.

The SHARE study had been designed by balanced randomisation and our analysis plan had to reflect this. We had recently developed methods that allowed a non-parametric analysis of a balanced randomisation to adjust for individual or group-level covariates (Randomization inference for balanced cluster-randomized trials Gillian M., Butcher, Isabella *Clinical Trials,* Volume 2, Number 2, April 2005, pp. 130-140(11)). Such analyses should use only a small number of group-level (here school-level) covariates. We selected one such measure taken from a factor analysis of several measures (called PRIN1) that had proved highly predictive of early sexual intercourse (Butcher I, PhD thesis, Napier University, 2005). Thus our analysis plan consisted of

- Comparing outcomes for pupils in SHARE and control schools with a restricted randomisation test
- Additionally, using the same test adjusted for social-class/ school leaver, cohort and for PRIN1
- Investigating the interactions of any effect of the intervention on outcome for 1) early school leavers 2) Social class (not early leavers only) and 3) cohort.

Finally, to comply with the consort guidelines for cluster trials we planned to carry out a model based analysis using a random-effects model so that we could quote ICC values in the paper. This was only possible for the binomial outcomes because we did not have any estimates of within group variances for the other measures.

All the main analyses were to be by intention to treat, ignoring how well or badly the SHARE program had been delivered. Such data were available from our process evaluation, and it was agreed that after our main analysis was complete we would carry out some on-treatment analyses. This was done by having the SHARE schools divided into three groups according to the quality of

their delivery of SHARE. This was done by someone who had not seen the ISD linked data. Exploratory analyses were then carried out to determine if this pattern could explain between school differences.

Appendix 1: Agreement with ISD Scotland

Ad hoc Number: **IR 2004-00891**

Study Title

Impact of SHARE sex education programme on pregnancies & terminations (miscarriage) at 5 years follow up: a randomised trial

Customer(s) Details

Name: Marion Henderson
Address: Social & Public Health Sciences Unit
University of Glasgow
4 Lilybank Gardens
GLASGOW
G12 8RZ

Tel: 0141 357 3949
Fax: 0141 337 2389

E-Mail: marion@msoc.mrc.gla.ac.uk

Aims

The aim of the study is to evaluate whether a school sex education programme (SHARE), designed according to the best educational theories and practices, and incorporating insights and theories from recent social science research on young people's sexual behaviour, has had any effect on young women's rate of births and therapeutic abortions.

Twenty-five secondary schools were recruited to this trial and randomly allocated to receive the SHARE intervention or be controls. Between 1996 and 1999 the two year SHARE programme was delivered to two successive cohorts of S3 and S4 pupils in the thirteen intervention schools. Pupils in all 25 schools have been followed up at age 16, 18 and now 20, using self-complete questionnaires. We now seek to establish the effect of the SHARE programme on the cumulative birth and therapeutic abortion rate by the age of 20. This will be the most valuable outcome measure since it is not subject to reporting bias or attrition.

Summary

To provide aggregate SMR 01 and SMR 02 data on therapeutic abortions, still births and live births by the age of 20 for the SHARE sample (identified by name, date of birth and postcode), broken down by school and cohort.

3.1.1

Cases

3.1.2 Cases comprise all young women who agreed to participate in the SHARE trial in two successive cohorts in 25 schools, recruited in S3 in 24 schools and in a further school in S5. Data were obtained through self-complete questionnaires first administered in classrooms under exam conditions.

Record Linkage within ISD

(i) Linkage of SHARE Females Cohort to SMR2 to December 2001

At present all records on the maternity and neonatal file are linked via the mother record, for example each year's SMR2 file is linked to the existing SMR2 records on the database. This provides a file with each mother's maternity records grouped together. SMR11 records and GRO Birth records are then linked to the SMR2 records, which provides the baby information for each pregnancy in the group.

(ii) Linkage of SHARE Females Cohort to Acute hospital discharges (SMR01) – data currently complete to December 2003

The linked data set required for this analysis contains linked SMR1/01, SMR6, SMR4/04 and Registrar General's death records. SMR1/01 (Scottish Morbidity Records 1) cover all non-obstetric and non-psychiatric discharges from NHS hospitals in Scotland. SMR6 are cancer registration records and SMR4/04 are mental health inpatient records. All patient records including deaths for each patient are linked together using 'probability matching'. The 'probability matching' algorithm uses all available identifying information (name, date of birth, postcode, hospital patient reference number etc.) to link the records.

Within these 'patient record sets', the SMR1/01 records are grouped into continuous stays. A continuous stay is a continuous period of time spent as an inpatient or day case in hospital regardless of any transfers between specialties or hospitals. For example, a patient may be admitted with an Acute Myocardial Infarction in a specialty of General Medicine, be transferred to Cardiology then transferred again to Geriatric Assessment before discharge. This single continuous stay would have generated three separate SMR1/01 discharge records which linkage can bring together.

This linked data set currently contains SMR1/01, cancer registration (SMR6), mental health (SMR4/04) and death records for the period 1981 onwards and holds data on over 5 million patients with over 20 million contacts within the acute hospital sector.

3.1.2.1 Accuracy of the data

In a world with perfect recording of identifying information and unchanging personal circumstances, all that would be necessary to link records would be the sorting of the records to be matched by personal identifiers. In the real world of data however, for each of the core items of identifying information used to link the records (surname, initial, year, month and day of birth), there may be a discrepancy rate of up to 5% in pairs of records belonging to the same person. Thus exact matching using these items could miss up to 25 % of true links.

To allow for the imperfections of the data, the linkage system uses methods of probability matching which have been developed and refined over the last thirty years. Despite the size of the data sets, linking the records consists of carrying out the same basic operation over and over again. This operation is the comparison of two records and the decision as to whether they belong to the same individual.

The linkage methodology is aimed at squeezing the maximum amount of discrimination from the available identifying information. Thus the distribution of probability scores differs for each kind of linkage.

Chief Scientist Office (CSO)

As part of this formal specification, a costing has been calculated on a cost recovery basis and is our best estimate of the resources required to fulfil the relevant project aims. Many research projects undertaken by ISD are supported through funding from a variety of bodies and researchers should notify ISD if existing funding is in place. Where funding remains to be pursued, one alternative is to utilise a service provided through partnership between ISD and the CSO. Since January 2001 two full time statisticians have been in post within ISD's Medical Record Linkage Team with a remit to undertake record linkage related projects. Researchers wishing to utilise this service should submit the relevant grant application form to the CSO, adding the estimated costs to any existing funding requirements.

Detailed information on the role of CSO and instructions on how to make an appropriate grant application can be obtained through their website.

http://www.show.scot.nhs.uk/cso/

Alternatively general enquiries should be directed to the following address.

Chief Scientist Office,
Scottish Executive Health Department,
St Andrew's House,
Regent Road,
Edinburgh, EH1 3DG

Tel: 0131 244 2248
Fax: 0131 244 2285

DATA SPECIFICATION

Please tick to indicate which Data schemes are involved:

SMR 00 Outpatient attendances	☐
SMR 01 General acute inpatient and day case discharges	☒
SMR 02 Maternity inpatient and day case discharges	☒
SMR 04 Psychiatric inpatient admissions, residents and discharges	☐
SMR 06 Scottish cancer registrations	☐
SMR 10 School health entrant and leaver records	☐
SMR 11 Neonatal discharges	☐
SMR 20 Scottish Cardiac Surgery Register	☐
SMR 50 Geriatric Long Stay	☐
RG Data Death Registrations	☐
RG Data Birth Registrations	☐

Other data sources

Survey questionnaires *Please specify the survey(s)*	☐	
Clinical trials	☐	
Health Board records	☐	
Hospital records	☐	
GP records	☐	
Employee's records	☐	
Other *Please specify*	☒	Approximately 4210 SHARE Females records
No other sources	☐	

Data Available for Linkage

In order to identify the records of the patients entered into the study we use probability matching on identifiable variables (the researchers have ten variables collected at interview):

First Name
Last Initial
Date of Birth
Postcode

Date of Delivery
Gender of Baby

A linkage program will be developed for this data set to maximise results. The results from the linkage process will be returned in an agreed format that meets the PAC agreement.

ESTIMATED RESURCES/COSTS

Data File

For the file of patients entered into the study we require a flat file with fixed length format (ASCII) containing the identifying data items.

Pre-processing

The file will be validated for linkage and the processing required prior to linkage will be added.

1 person day @ £200 per day £200.00

Linkage Program

A linkage program will be developed for this data set and tested to maximise results. Using probability matching theory we will develop a linkage program which calculates odds and matches the external records to our databases (including SMR02, SMR04, RG Deaths, RG Births and SSBID).

The results from the linkage process will be returned in an agreed format.

- Linkage of SHARE Females Cohort to SMR2

  10 person days @ £200 per day £2000.00

- Linkage of SHARE Females cohort to SMR01

  7 person days @ £200 per day £1400.00

3.2 Summary of total days/costs required

Including pre-processing of hospital records, linkage programs and analytical requirements.

3.2.1.1.1.1 Description	3.2.1.1.2 No. days	3.2.1.1.3 Cost
Pre-processing	1 person day @ £200	£200.00
Linkage programmes: A linkage program will be developed for this data set and tested to maximise results – SMR2	10 person days @ £200	£2000.00
Linkage programmes: A linkage program will be developed for this data set and tested to maximise results – SMR1	7 person days @ £200	£1400.00
Analytical resources: Results will be output in an agreed format.	1 person days @ £200	£200.00
Total	19 days	£3800.00

Appendix 2: Letter to ISD specifying details of linkage

MRC
Medical Research Council

Social and Public Health Sciences Unit
University of Glasgow
4 Lilybank Gardens
Glasgow
G12 8RZ

Telephone +44 (0) 141-357 3949
Fax +44 (0) 141-337 2389
Web www.msoc-mrc.gla.ac.uk
E-mail Marion@msoc.mrc.gla.ac.uk

Mr James Boyd and Mr Alan Finlayson
Information & Statistics Division
1st Floor
Area 122A
Gyle Square
1 South Gyle Crescent
Edinburgh
EH12 9EP

05 August 2004

Dear Mr James Boyd and Mr Alan Finlayson,

3.3 Impact of SHARE sex education programme on pregnancies & terminations (miscarriage) at five year follow up: a randomised trial

At last, I now have as complete a data set as possible for linkage of the SHARE females! Overleaf, I have provided information on the three different aggregations that we would appreciate you providing us with information. I have also provided you with information on all the variables that I have provided for linkage purposes (see CD 'ISD.SAV' (SPSS format as okayed by Alan. Also, this letter is on the CD – 'ISDlet.doc').

As discussed at our meeting, James, what would be most helpful to us is for each group (in each of the three different aggregation strategies (see ISD1, ISD2 & ISD3 on data set and explanation below) to know:

- **Each event of pregnancy**;
- **What was the outcome of each pregnancy**… i.e. termination, miscarriage or live birth;
- **What age the woman** was at the time of the event;
- **Should one woman within a group have more than one event could that be flagged** (e.g. woman X pregnancy at 16 terminated, same woman X pregnancy at 18 live birth)… this is because it could affect the outcome of our trial whether two or three *different* woman within a group all have pregnancies compared with one woman having two or three pregnancies

It would be very helpful to know how long you anticipate the linkage taking and how much it is likely to cost.

Please do not hesitate to contact me should you wish to discuss anything.

Yours sincerely,

Marion Henderson
Senior Scientific Officer

Descriptions of variables included in data set ISD.SAV (ready only)

IDNO = is our project's unique identification number for participants

ISD1 = Linkage variable 1, groups Non leavers by social class, school, cohort, arm of trial

ISD2 = Linkage variable 2, groups leavers by social class, school, cohort, arm of trial

ISD3 = Linkage variable 3, groups pupils that have poor attendance by cohort and arm of trial

FORNAME

SURNAME

SEXF = all are female

DAY = day month year for date of birth
MONTH
YEAR (NB where date of birth is missing then year of birth will equal 9). To facilitate linkage for those with missing date of birth, the participants were all born between 1981 to 1985 inclusive.

PCBEST = the postcode's originally provided by participants

CPC1 = change of address postcode 1
CPC2 = change of address postcode 2
CPC3 = change of address postcode 3
SCHOOLPC = school postcode for those with postcode data missing (small minority)

This document contains three tables (the 1st is long… ISD2 & ISD 3 are on the last two pages).
Table 1 below indicates the number of young women falling into each category of interest to the SHARE study. The variable ISD1 contains the aggregate groupings (categories) of interest.

Decoder for ISD1:

Units represent cohort	0 = cohort 1 & 1 = cohort 2
Tens represent early school leavers	0= non-leavers & 10 = leavers
Hundreds represent non leavers social occupation	0 = non-manual, 100 = manual & 800 = leavers
Thousands represent school attended (1000 to 25000)	

Table 1 ISD1

ISD1 (variable that represents the aggregate groupings (categories) of young women in the SHARE sample)	Frequency (number in each group of interest)
1000	50
1001	29
1100	23
1101	29
1810	5
1811	20
2000	9
2001	6
2100	21
2101	9
2810	4
2811	3
3000	10
3001	7
3100	23
3101	20
3810	16
3811	21
4000	53
4001	53
4100	18
4101	21
4810	9
4811	17
5000	30
5001	23
5100	35
5101	31
5810	32
5811	44
6000	12
6001	10
6100	58
6101	43
6810	25
6811	15
7000	33
7001	32
7100	29
7101	25
7810	23

ISD1 (variable that represents the aggregate groupings (categories) of young women in the SHARE sample)	Frequency (number in each group of interest)
7811	20
8000	17
8001	18
8100	15
8101	31
8810	35
8811	27
9000	16
9001	11
9100	35
9101	39
9810	23
9811	24
10000	66
10001	48
10100	36
10101	26
10810	12
10811	18
11000	28
11001	40
11100	30
11101	18
11810	8
11811	13
12000	7
12001	18
12100	34
12101	54
12810	28
12811	13
13000	52
13001	60
13100	47
13101	57
13810	19
13811	24
14000	44
14001	54
14100	21
14101	16
14810	15

ISD1 (variable that represents the aggregate groupings (categories) of young women in the SHARE sample)	Frequency (number in each group of interest)
14811	10
15000	57
15001	60
15100	22
15101	21
15810	13
15811	11
16000	26
16001	25
16100	32
16101	40
16810	34
16811	27
17000	21
17001	20
17100	25
17101	18
17810	13
17811	13
18000	45
18001	45
18100	35
18101	18
18810	16
18811	29
19000	16
19001	19
19100	27
19101	30
19810	32
19811	21
20000	14
20001	23
20100	36
20101	34
20810	42
20811	35
21000	58
21001	48
21100	54
21101	47
21810	31

ISD1 (variable that represents the aggregate groupings (categories) of young women in the SHARE sample)	Frequency (number in each group of interest)
21811	18
22000	25
22001	12
22100	42
22101	38
22810	14
22811	18
23000	41
23001	39
23100	14
23101	29
23810	13
23811	14
24000	19
24001	22
24100	46
24101	50
24810	23
24811	34
25000	9
25001	9
25100	19
25101	38
25810	30
25811	17
Total	4069

Decoder for ISD2:

Units represent cohort 0 = cohort 1 & 1 = cohort 2
Tens represent early school leavers 0= non-leavers & 10 = leavers
Hundreds represent non leavers social occupation 0 = non-manual, 100 = manual & 800 =NON-leavers
Thousands represent arm of the trial 1000 = intervention & 8000 = control

Table 2 ISD2

ISD 2 (variable that represents the aggregate groupings (categories) of young women in the SHARE sample)	Frequency (number in each group of interest)
1010.00	55
1011.00	59
1110.00	184
1111.00	200
1800.00	787
1801.00	758
8010.00	73
8011.00	78
8110.00	203
8111.00	169
8800.00	748
8801.00	755
Total	4069

Decoder for ISD3:

Units represent low attendance / no SHARE contact 0 = contact & 1 = no contact
Tens represent cohort 0 = cohort 1 & 1 = cohort 2
Hundreds represent arm of the trial 0= control & 10 = intervention

Table 3 ISD3

ISD 2 (variable that represents young people with little exposure to SHARE or control equivalent)	Frequency (number in each group of interest)
0	1009
1	51
10	981
11	40
100	1015
101	44
110	989
111	81