Scottish Health Survey

'09

User Guide

A survey carried out on behalf of The Scottish Government Health Directorates and NHS Health Scotland

Scottish Centre for Social Research

Department of Epidemiology and Public Health, University College London

& The MRC Social and Public Health Sciences Unit, Glasgow

1. Background

The data files contain data from Scottish Health Survey 2009 (SHeS09), the fourth of a series of surveys designed to monitor trends in the nation's health. The Scottish Health Survey was commissioned by the Scottish Government and carried out by the Scottish Centre for Social Research*,* the Department of Epidemiology and Public Health at University College, London, and the CSO/MRC Social and Public Health Sciences Unit, Glasgow.

The aims of the Scottish Health Survey series are:
- to provide data about the nation's health;
- to estimate the prevalence of particular health conditions;
- to estimate the prevalence of risk factors associated with these conditions;
- to examine differences between population subgroups; and between Scotland and England
- to contribute towards monitoring progress towards selected health targets;
- to monitor trends in the population's health over time

This user guide provides an overview of the study and the key elements of the dataset that all users need to know before conducting any analysis. It is designed to be read in conjunction with the documentation supplied with the dataset and the Technical Report to the 2009 Scottish Health Survey Report which is available online here.

2. Survey Design

The Scottish Health Survey was designed to provide data at a national level about the population living in private households in Scotland. Although it wasn't designed to provide sub-national regional data on an annual basis, Section 3 describes the potential for regional analysis that will exist in future years at Health Board level. The sample for the 2009 survey, as in previous years, was drawn from the Postcode Address File (PAF). An initial sample of 12,668 addresses were selected and grouped into 503 interviewer batches, with around 45 batches covered each month between January and December 2009. The addresses were comprised of three sample types:

- 7,588 formed the main sample, at which adults and children were eligible to be selected for interview;
- 4,312 addresses formed an additional child boost sample at which only households containing children aged 0-15 were eligible to participate;
- 768 addresses (in Grampian, Fife and Borders) formed the Health Board boost sample, at which only adults were eligible for interview.

All private households in the sample were eligible for inclusion in the survey (up to a maximum of three households per address).

Information was obtained directly from persons aged 13 and over. Information about children under 13 was obtained from a parent or guardian with the child present.

Data collection involved a Stage 1 interview, and if applicable, adults in the main sample also had a follow-up visit from a specially trained nurse. Of the 7,588 main addresses issued, 2,152 were flagged as the 'nurse sample'. At these addresses all adults (16+) interviewed at Stage 1 were eligible to take part in the Stage 2 follow-up nurse visit. There were no nurse visits at the remaining addresses or for the child boost or health board boost samples.

The contents of the Stage 1 interview and the nurse visit are listed in Appendix A.

Interviewing was conducted throughout the year to take account of seasonal differences.

3. Key changes to the survey in 2009

A number of changes to the survey methodology were introduced from 2008 onwards. The implications of these for data users are set out below:

Move to a continuous format - Prior to 2008, the survey had been carried out on three separate occasions (1995, 1998 and 2003). It is now being carried out continuously between 2008 and 2011. The majority of the questionnaire is going to be the same in every year, with a smaller number of topics included in two of the four years (2008 and 2010, or 2009 and 2011). This has implications for the topic coverage and the sample size, as set out below.

Core and modular questionnaire structure: Adults –The Stage 1 interviewer visit for adults now has a core and modular structure with a core set of questions asked of the whole sample and two modules of questions which are only asked of a proportion of the sample. Addresses are allocated at random to one of two modules (version A or B) before they are assigned to interviewers; participants answer questions from the core and from one of the two modules. Core questions will be included in the survey every year and these will be analysable by Health Board after four years. For some larger Health Boards analysis will be possible sooner than this, a paper setting out the various regional analysis options is available from the Scottish Government website here. Version A of the two modules asked of a sub-sample is the 'rotating' biennial module. In 2009 it contained a range of questions on accidents, dental services, discrimination and harassment, and stress at work. These questions will be asked again in 2011. A module with different topics was included in 2008 and will be repeated in 2010. A specific weight is provided for analysis of Version A data. The variable labels for all the items in this version start with "VERA" to help users identify which weight to use.

Questionnaire structure: Children – The Stage 1 interviewer visit for children does not have a core and modular structure so the questions are the same every year. The child sample is designed to be representative at the national level, regional analysis is not possible.

Knowledge, Attitudes and Motivations module (KAM) – Version B of the two modules asked of a sub-sample is known as the KAM module and is a replacement for the Health Education Population Survey (HEPS) which was previously run by NHS Health Scotland. One adult per household is selected at random to answer the KAM module. A specific KAM weight is provided for analysis of KAM data. The variable labels for all the items in the KAM module start with "KAM". When the 2008 SHeS dataset was deposited the KAM data was not yet available for public use. However it is included in the 2009 dataset and a version of the 2008 datafile with the KAM data included is now available.

Reduced Stage 2 nurse visit – Unlike the three previous surveys, only a sub-sample of adults are eligible to take part in the nurse visit in the 2008 -11 surveys. No nurse visits are carried out with children, or with adults in the Health Board boost sample. The nurse sample is designed to allow for analysis of nurse data at the national level after one year, but further disaggregation by age and gender will be limited by the sample size. More detailed analysis of sub-groups within the sample will be possible once two or more years of data have been aggregated.

Unclustered sample design – Clustered sample designs increase the standard errors around estimates in population surveys. The sampling was constructed so that each year's sample was clustered but the four-year sample 2008-2011 will be unclustered. The Technical Report contains a more detailed description of the sample design. The main

reporting outputs from the survey published by the Scottish Government take the clustering into account when calculating standard errors, confidence intervals and significance levels. Data users who do not adjust for the sample design in their analysis will not be able to reproduce the standard errors or confidence intervals presented in the reports. The methods for doing this are complex but packages such as Stata, SAS and some versions of SPSS can do this adjustment. For advice on this contact the Scottish Government Health Survey team directly on this: scottishhealthsurvey@scotland.gsi.gov.uk.

Optional NHS Health Board boost – NHS Health Boards were given the option to boost their samples beyond the level which is funded centrally. In 2009, Borders, Fife and Grampian Health Boards chose to boost the number of adults interviewed in their areas. These interviews only included the core questions (described above) and did not involve nurse visits. These cases have been included in the main 2009 file and their additional selection probability has been adjusted for in the weighting scheme (see below). As a result of these boosts the data in Borders and Fife will be analysable after two years rather than four. The boost in Grampian was intended to produce data suitable for analysis after one year. However, the response rate was lower than anticipated so there are slightly fewer cases than the recommended minimum sample size for Health Board analysis.

4. Documentation

The documentation has been organised into the following sections
- Interview (contains the CAPI documentation for household and individual questionnaires, nurse visit questionnaires, self-completion booklets and showcards)
- Data (contains the list of variables in the file, a list of derived variables, and a list of variables used in the main 2009 report tables)
- Other instructions (contains interviewer, nurse and coding & editing instructions)

5. Using the data

The 2009 data consists of one individual level file and one household level file:

SHeS09i.sav	10138 records	contains data for all individuals who gave an interview. It contains information from the household questionnaire, main individual schedule, self-completions, the nurse visit (where one occurred) and the KAM interview (if the respondent had been selected for this part of the study).

SHeS09h.sav	13400 records	contains data on household, and sex and age of all individuals in co-operating households.

5.1 Variables on the files

Each of the data files contain questionnaire variables (excluding variables used for administrative purposes and any variables that are potentially disclosive) and derived variables. The variables included in the individual file are detailed in the "**Variables**" document in the data section of the documentation. This document is the best place to look at in order to plan your analysis. It includes:
- Major categories of variables (eg Smoking, Anthropometric measurements)
- Sub categories of variables (eg Nicotine replacements within the Smoking category)
- Source of each variable (eg Individual questionnaire, Nurse visit, Derived variable etc.)

Once you have decided which variables to include in your analysis, you should look up details of the question wording using the interview section documentation (all variables on the data file

are given by name in the copy of the interview schedules provided), or use the "**Derived Variables**" document in the data section of the documentation for the syntax which produced the derived variables. You cannot rely on the individual variable and value labels to always capture the detail of the question asked, or the answer categories used, so reading the interview documentation is essential.

To assist users, particularly those unfamiliar with the survey series, we have produced a guide to the variables used in the tables in the main 2009 report. In most cases this identifies the key variables for each of the main topics covered in the survey.

Existing questions that changed notably between survey years, for example changes to their wording or response categories, have been given new variable names including the figure "08". In most cases these have also been highlighted in the interview documentation.

5.2 Weighting variables

Weighting has been used to correct for different selection probabilities and for non-response. The non-response weights were designed to adjust for non-contact and for refusals of entire households, the non-response of individuals within responding households, and non-response to specific aspects of the study (the KAM module, the nurse visit, the blood sample). Separate weights exist for adults and children. The aim of each set of weights is that each of the main datasets (households, individuals responding to the different parts of stage 1, individuals who see a nurse, and individuals who give blood) can be treated as broadly representative of the general household population. The weights were designed so that the weighted age/sex profile of the sample matched the GROS 2009 mid-year household population estimates for Scotland.

Weight name	When it should be used	Primary aim of this weight
HH09wt (household)	This weight is used to derive the individual weights (below) and is not for analysis purposes. It is available in the household data file only	Adjusts for differences between responding and non-responding households
Int09wt (adults)	Analysis of items in core questionnaire	Adjusts for differences between responding and non-responding individuals within cooperating households
VerA09Wt (adults)	Analysis of items in version A of questionnaire (variables with labels starting "VERA")	Adjusts for the fact that the Health Board boost sample interview didn't include the version A module (Int08wt included the heath board boost cases)
KAM09wt (adults)	KAM module analysis (variables with labels starting "KAM")	Adjusts for differences between those who answered the KAM module and those who did not Adjusts for the selection of one adult per household
Cint09wt (children)[1]	Analysis of child data	Adjusts for the selection of just two children per household Adjusts for differences between responding and non-responding households Individual non-response

[1] Adults have values of 0 for cint08wt. Children aged 0-15 have values of 0 for all the adult weights.

		weighting was not needed as response was high among children in cooperating households
Nurs09wt (adults)	Nurse questionnaire analysis (with exception of blood samples)	Adjusts for differences between adults who had, and those who did not have, a nurse visit
Blod09wt (adults)[2]	Blood sample analysis	Adjusts for differences between those who gave a blood sample and those who did not

The different aspects of the survey are obviously all linked. **The weighting variable selected should always match the dependent variable in the analysis**. For example, to analyse data collected in the nurse interview on its own, or in relation to data collected in the core interview questionnaire (e.g. cotinine level by smoking history, lung-function by respiratory symptoms) the nurs09wt weight should be used. To look at the blood data on its own, or by other nurse data, or by interview data in the core interview questionnaire, the blod08wt should be used. Conversely, to look at data in the core interview questionnaire by nurse or blood data (e.g. smoking history by cotinine level, or respiratory symptoms by lung-function, or CVD prevalence by total cholesterol level) the int09wt weight should be used.

Full details of the weighting are provided in the main Technical Report.

5.3 Multicoded questions

Some questions in the survey enabled participants to give more than one answer. In the final dataset each of the answer options has been converted into a binary variable with the people who selected that option coded 1 and the rest coded 0.

As an example, question CONSUBX in the adult nurse interview is a "CODE ALL THAT APPLY" question which asks "Have you eaten, smoked, drunk alcohol or done any vigorous exercise in the past 30 minutes ?". The code frame consists of five values:

    1 -   eaten
    2 -   smoked
    3 -   drunk alcohol
    4 -   done vigorous exercise
    5 -   none of these

The five answer options have been converted into five separate binary variables as follows:

CONSUBX1 - code 1: those who ate in the last half hour; code 0: those who didn't.
CONSUBX2 - code 1: those who smoked in the last half hour; code 0: those who didn't.
CONSUBX3 - code 1: those who drank alcohol in the last half hour; code 0: those who didn't.
CONSUBX4 - code 1: those who did vigorous exercise in the last half hour; code 0: those who didn't.
CONSUBX5 - code 1: those who did none of the above in the last half hour; code 0: everyone else.

Because a respondent could have replied with more than one answer, that respondent could have a value 1 for a number of these variables (however, the nature of the question dictates

---

[2] Adults who responded to both the main interview and nurse visit but who did not give a blood sample have a value of 0 for **blod08wt**. This means that some functions in SPSS will generate an error message stating that the weight used contains negative or zero values. This can be ignored.

that having a code 1 at CONSUBX5 precludes having a code 1 at any of the variables CONSUBX1 – CONSUBX4). The missing values are the same across all five variables.

5.4 Missing values conventions

-1   Not applicable: Used to signify that a particular variable did not apply to a given respondent usually because of internal routing. For example, men in women only questions.

-2   Schedule not applicable: Used for variables when the respondent was not of the given age range or sample type.

-6   Schedule not obtained: Used only for variables on the self-completion schedules this code indicates that a self-completion booklet was not completed for this respondent

-8   Don't know, Can't say.

-9   No answer/ Refused

These conventions have also been applied to most of the derived variables but the derived variable specifications should be consulted for full details.

5.5 Valid cases

In the 2009 Scottish Health Survey report, as in previous reports, cases were excluded from the analysis of anthropometric and other physiological measurements if their measurement was invalid. For example, those who had smoked, drunk or eaten within 30 minutes of having their blood pressure taken were excluded from the main blood pressure variables as this can affect blood pressure. The Variables document gives details of which variables show only valid codes, and which also include invalid answers.

5.6 Derived variables

In addition to the questions and measurements collected directed in the survey, a large number of derived variables have been created for use in the analysis. These variables are sometimes just straightforward recodes of existing variables, for example a summary variable that collapses some categories to make the data more succinct, or a conversion of continuous data (e.g. age) into categories (e.g. age groups). In most cases the derived variables make use of the underlying data in a number of variables to create new variables. For example, the height and weight data is used in combination to derive the Body Mass Index variable. It can sometimes appear to users that there are multiple measures of the same item within the dataset, especially in more complex parts of the questionnaire (e.g. the smoking and drinking sections). In these instances it is advisable to use the derived variable listing provided, or the listing of variables used in the report, to identify variables for potential analysis, and to refer back to the questionnaires to confirm your selection, rather than to look at the questionnaire documentation first. All derived variable labels start with "(D)" to help distinguish them from other types of variable. Some of the more complex derived variables require the use of look-up tables (e.g. children's BMI groups) and the syntax has not been included. Further information on these variables can be obtained from research team.

5.7 Socio-economic classification and social class measures
As detailed in the main technical report, the survey uses the National Statistics Socio-Economic Classification (NS-SEC) introduced in 2001. Information about all adult respondents' employment history is collected and where applicable two NS-SEC codes are derived: one for the Household Reference Person (HRP) and one for the individual respondent. The HRP is the

householder with the highest income within the household. In addition, adult respondents were asked what their mother and father did for a living when the respondent was 14 so there are also additional NS-SEC codes for one or both parents (if sufficient information was collected).

The NS-SEC was a replacement for the Registrar General's Social Class classification system. Although this classification system has been discontinued it is possible to derive synthetic versions of it based on the NS-SEC coding. These may be helpful to users interested in time series analysis based on the RG codes used in the 1995 and 1998 surveys. As with NS-SEC, the RG social class codes are available for the HRP and the individual.

	Variable name prefixes	
Classification level	**NS-SEC**	**RG Social Class**
Household reference person	HPNSSEC	SCHRP
Individual respondent	NSSEC	SCALL
Respondent's father (when R aged 14)	FANSSEC	N/A
Respondent's mother (when R aged 14)	MANSSEC	N/A
Respondent's parent (when R aged 14) – highest of father/mother	PNSSEC	N/A

More information about NS-SEC and RG social class is available from the ONS website here.

6. SHeS 2009 Report

Further information about the Scottish Health Survey 2009 is available in:

Bromley C, Given L, and Ormston R (eds) *The Scottish Health Survey 2009,* Edinburgh: Scottish Government, 2010.

The full report, and a set of further tables with selected results for topics not covered in the report, is available on the web at:

http://www.scotland.gov.uk/Topics/Statistics/Browse/Health/scottish-health-survey/Publications

Users might also be interested in viewing the rest of the Scottish Health Survey website:

http://www.scotland.gov.uk/Topics/Statistics/Browse/Health/scottish-health-survey

This contains a large amount of useful information including the background to the study, the consultation that informed the changes made to the survey in 2008, and plans for future dissemination.

APPENDIX A
SCOTTISH HEALTH SURVEY 2009 – CONTENTS

1) Interviewer visit

Questionnaire Items	Respondent	Version
Household information	Household reference person/spouse	Core
General health	All ages	Core
Cardio-vascular disease and use of health services	16+	Core
Accidents	All ages	A
Adult physical activity	16+	Core
Child physical activity	2-15 years	Core
Child eating habits	2-15 years	Core
Fruit and vegetable consumption	2+	Core
Smoking and drinking	16+*	Core
Dental health	16+	Core
Dental services	16+	A
Social capital	16+	A
Discrimination and harassment	16+	A
Employment status, educational background	16+	Core
Stress at work	16+	A
Ethnic background	All ages	Core
Religion	16+	Core
Parental social class & health history	16+	Core
Height measurement	2+	Core
Weight measurement	All ages	Core
Knowledge, Attitudes and Motivations module	16+ (one adult selected at random)	B
Self completions		
Smoking	16-17*	Core
Drinking	16-17*	Core
General Health Questionnaire 12	13+	Core
Warwick-Edinburgh Mental Wellbeing Scale	16+	Core
Contraceptive use	16+	Core
Strength and difficulties questionnaire	Parents of 4-12 year olds	Core
Attitudes to health CASI self-completion**	16+ (one adult selected at random)	B

*18 - 19 year olds answer these questions either in the self-completion or in the main interview, at the interviewers' discretion.

2) Nurse Visit

NURSE Measurements	Respondent
Details of prescribed drugs	16+
Vitamin supplements	16+
Stress, anxiety and self-harm	16+
Food poisoning	16+
Nicotine replacement products	16+
Blood pressure	16+
Waist and hip circumferences	16+
Demi-Span	65+
Lung Function	16+
Blood sample	16+
Saliva sample (for cotinine)	16+
Urine sample (for dietary sodium)	16+