DATA PAPER

Epidemiology Data from the Scottish Health and Ethnicity Linkage Study (SHELS)

Judith Fernandez[1], Anne Douglas[1], Genevieve Cezard[1] and Raj Bhopal[1]

[1] Scottish Health and Ethnicity Study, University of Edinburgh, UK

We linked the 2001 Scottish Census, which contains ethnicity, socio-economic and demographic data to health and death records, creating an anonymised retrospective cohort study of 4.65 million people to assess the association between ethnicity and health outcomes in Scotland. The databases contain data mostly from hospital discharge and mortality records, but also from other registers. The databases are stored in a safe haven at the National Records of Scotland (NRS). NRS is currently exploring the feasibility of making Scottish Health and Ethnicity Linkage Study (SHELS) data open access while ensuring that the same level of confidentiality is maintained. If SHELS becomes open access it could be reused, with the appropriate approvals, to assess the influence of other socio-economic or demographic measures on the Scottish population's health.

Overview

Spatial coverage
The data were collected for all Scotland.

Temporal coverage

- May 1st 2001 to April 30th 2008 for cardiovascular diseases, cancer, breast cancer screening, mental health and maternal and child health.
- May 1st 2001 to April 30th 2010 for respiratory diseases, gastrointestinal diseases and primary care (pilot)
- May 1st 2001 to April 30th 2013 for the ongoing phase which aims to continue methodological work on the use of cardiovascular and respiratory risk factors from primary care and in addition, to study:
  - all-cause mortality
  - all-cause hospitalisation
  - hospitalisation and mortality for:
    - infectious and parasitic diseases
    - injuries, accident and poisoning
  - Uptake of bowel cancer screening and pathology of screen–detected bowel cancers

Species
Human population

Methods

Steps
The methods and many of the findings of our retrospective cohort study (SHELS) have been published [1-12]. We followed a strict protocol that preserved anonymity and maintained separation of personal data from the Census and clinical data. **Figure 1** shows our approach to linkage. We used computerised probability matching of names, sex, addresses and dates of birth to link the 2001 census for Scotland, to the Scottish Community Health Index (CHI), which is a register of patients using the NHS (National Health Service). At this stage, other data fields in the two datasets were excluded. CHI and the census unique numbers were encrypted. A one-way cryptographic ('hashing') algorithm was used to encrypt the CHI number. The census number was encrypted using an algorithm developed by the National Record of Scotland (NRS).

About 95% (approx 4.65 million) of the people participating in the 2001 census (4.9 million) were linked as above to the Scottish Community Health Index, with 85% or more linked in every ethnic group. This represents about 92% of the 2001 population. This linked file of encrypted CHI and census numbers is the key to subsequent linkage of any health data to the 2001 census records.
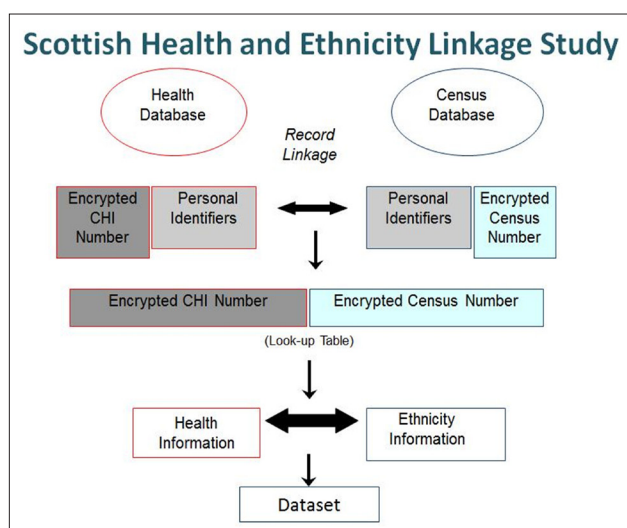
Figure 1: Overview of record linkage process.

Using our retrospective cohort, we were able to analyse ethnic variations in various health and healthcare areas: cardiovascular diseases [2-6], cancer [7-9], maternal and child health [10], mental health [11], gastrointestinal diseases and respiratory diseases (to be submitted for publication in 2014). We have also linked primary care records from 10 general practices in Edinburgh and Glasgow.

Both hospitalisation diagnoses and causes of death (see **tables 2** and **3** in appendix) were available in each health area dataset. Other health datasets were linked for the analyses of maternal and child health and mental health outcomes (see **table 1** in appendix).

These morbidity and mortality data were examined in relation to ethnicity, adjusting for demographic and socio-economic measures obtained from the 2001 census (See **table 1** in appendix).

Sampling strategy
Not relevant here.

Quality Control

1. The health data undergo quality control procedures by Information Services Division (ISD) of NHS National Services of Scotland. A Data Quality Assurance team ensure that the ISD health records are accurate, consistent and comparable across time and between sources. (http://www.isdscotland.org/Products-and-Services/Data-Quality/ http://www.isdscotland.org/Products-and-Services/Data-Quality/ – accessed 17/09/2014, contact: nss.isd-dmdataquality@nhs.net)
2. The Census data have undergone quality control procedures by NRS. (http://www.gro-scotland.gov.uk/files2/the-census/2001-census/census-assessment.pdf – accessed 17/09/2014, contact: customer@gro-scotland.gsi.gov.uk)
3. Extracted datasets were checked and incidence figures calculated to compare to official published statistics.

Constraints
95% of the Census was linked to the CHI and at least 85% linkage was achieved for each ethnic group. In any health dataset, a small percentage of records were not linked for reasons including:

a) CHI numbers not available or not linking to look up table
b) A person was not in Scotland in 2001
c) A person was not in the 2001 census

Privacy
Analytical datasets contain no personal identifiers. Statistical output is subject to a NRS disclosure protocol, and scrutiny by a disclosure committee. Researchers require government baseline security clearance for access to the data in a safe setting at NRS, as well as research governance training.

Ethics
The work was approved by the Multicentre Research Ethics Committee for Scotland and the Privacy Advisory Committee of NHS National Services Scotland, plus Community Health Index Advisory Group and Caldicott Guardian approval, where required.

Dataset description
Object name
SHELS data is not yet open access, for further information about the datasets see the tables in appendix. . A detailed metadata and data dictionary will be produced for each health extract once open access approval is agreed.

Data type
Secondary data, processed data.

Ontologies
None

Format names and versions
The datasets are stored in the safe haven as SAS dataset (.sas7bdat).

Creation dates
Creation date by health areas:

- Cardiovascular disease: April 2009
- Maternal and child health: October 2009
- Mental health: December 2009
- Cancer (including breast cancer screening): May 2010
- Gastrointestinal diseases: June 2012
- Respiratory diseases: November 2012
- Primary care risk factors: April 2013
- All-cause mortality: 2014 (ongoing)
- All-cause hospitalisation: 2014 (ongoing)
- Infectious and parasitic disease (including blood borne viruses): 2014 (ongoing)
- Injuries, accident and poisoning: 2014 (ongoing)

· Uptake of bowel cancer screening and pathology of screen–detected bowel cancers: 2014 (ongoing)

Dataset creators

A person independent of the core data analysis team linked the 2001 census numbers held by NRS to the CHI held by ISD, creating the SHELS look-up table with encrypted numbers. ISD extracted health datasets and NRS staff linked them to the Census using the look-up table held at NRS. Then, once the anonymised linked data-sets were created, the SHELS researchers were responsible for the data analysis.

Language

English

Programming language

SAS (alternatively STATA) code was used to prepare and analyse the data. Programs are available upon request.

Licence

Not open licence

Accessibility criteria

The linked anonymised Census data and the health data are accessible on a stand-alone computer in a locked room at NRS. Currently, access is restricted to SHELS researchers who have security clearance.

Repository location

Not accessible for the moment, please contact the Principal Investigator (raj.bhopal@ed.ac.uk).

Publication date

Not applicable

Reuse potential

The datasets contain health and socioeconomic data for a wide range of health areas. NRS is currently assessing the feasibility of making SHELS data open access and potentially this could be reused by other researchers, with appropriate agreements from ethics and privacy advisory committees, to analyse the association between health in the Scottish population, as recorded in datasets with CHI number and any variable in the Census.

Appendix

Health information	Census information
Dataset 1: CHD/Stroke	
- In patient and day-case discharge from the Scottish Morbidity Records SMR01 (general) database	
- Mortality data from NRS	
Dataset 2: cancers	- Ethnic group
- In patient and day-case discharge from SMR01 and SMR06 (cancer registrations) databases	- Religion, current
- Mortality data from NRS	- Religion of upbringing
- Screening data from the Scottish Breast Screening Programme	- Country of birth
Dataset 3: maternal and child health	- Age
- In patient and day-case discharge from SMR01 and SMR02 (maternity) database	- Sex
- Child health surveillance data	- Long term illness
- Scottish Birth Record data from NRS	- Self-assessed health
- Mortality data from NRS	
Dataset 4: mental health	- Marital status
- In patient and day-case discharge from SMR01 and SMR04 (psychiatric) databases	- Labour force status
- Mortality data from NRS	- Socioeconomic status
- Data from the Mental Welfare Commission for Scotland	- Highest qualification
Dataset 5: gastrointestinal	- Scottish Index of Multiple Deprivation decile
- In patient and day-case discharge from SMR01 (general) database	
- Mortality data from NRS	
Dataset 6: respiratory	- Car ownership
- In patient and day-case discharge from SMR01 (general) database	- Housing tenure
- Mortality data from NRS	- Household size
Dataset 7: all-cause hospitalisation*	- Numbers of rooms
- In patient and day-case discharge from SMR01 (general) database	- Urban/rural indicator
- Mortality data from NRS	
Dataset 8: all-cause mortality*	- Health board (Glasgow, Lothian, Tayside, Other)
- Mortality data from NRS	
Dataset 9: unintentional injuries and poisoning:*	
- In patient and day-case discharge from SMR01 (general) database	- Mobile (temporary) accommodation
- Mortality data from NRS	

Contd.

Health information	Census information
Dataset 10: all infectious and parasitic diseases*	
- In patient and day-case discharge SMR01 database	- Self-contained accommodation
- Mortality data from NRS	
Dataset 11: hepatitis C, hepatitis B and HIV*	- Central heating
- Records from the hepatitis C, hepatitis B and HIV datasets held by Health Protection Scotland	- Moved within last year
Dataset 12: bowel cancer screening*	- Economic activity last week
- Bowel Cancer Screening Key Performance Indicator (KPI) dataset held by ISD.	
- Cancer registrations SMR06 database	
Dataset 13: Primary care (sample of 10 GP practices)	
- Cardiovascular and respiratory risk	
- Further linked to Dataset 1, 6 and 8	
- Limited data on morbidity and prescribed drugs relating to cardio-vascular and respiratory diseases	

* Under preparation

Table 1: SHELS health and census datasets (with outcomes between 2001 and 2013, minimum follow-up of 7 years).

Phase	Health area	Health outcomes
Phase 1/2	Cardiovascular diseases:	- First myocardial infarction and survival (28-day) - First chest pain - First stroke - First episode of heart failure
	Mental Health	- First psychiatric disorder (any diagnosis) - First mood disorder - First psychotic disorder - Emergency detention certificate - Short-term detention certificate - Compulsory treatment order
	Cancer	- Any first cancer (excluding non-melanoma skin cancer) - First lung cancer - First prostate cancer - First breast cancer - First colorectal cancer - Breast cancer screening non-attendance
	Maternal and child health	- Mean maternal age - Mean gestational age - Birthweight - Proportion (%) of smokers (smoking history and smoking during pregnancy) - Proportion of preterm birth - Caesarean delivery - Proportion of breast feeding at 6-8 weeks
Phase 3	Respiratory diseases	- All-cause non cancer respiratory diseases - Asthma - COPD - All upper respiratory infection - Tonsillitis - All lower respiratory infection - Pneumonia - Influenza
	Gastrointestinal diseases	- Peptic ulcer disease - Oesophagitis - Gastritis - Gallstones - Pancreatitis - Irritable bowel syndrome - Appendicitis - Ulcerative colitis - Crohn's disease - Any inflammatory bowel disease - Diverticular disease
	Primary care	- Pilot study: data extraction and data quality assessment

Contd.

Phase	Health area	Health outcomes
Phase 4	Infectious diseases	- Mortality/hospitalisation for infectious or parasitic diseases and incidence/prevalence for HIV and hepatitis C and B
	Non intentional accidents	- Mortality/hospitalisation for accidents, injuries and poisonings
	Cancer	- Participation in and outcome of the Scottish Bowel Cancer Screening Programme
	Other	- All-mortality - All hospitalisation, length of stay and readmission - Utilisation of morbidity and risk-factor data from primary care

Table 2: SHELS health outcomes by health area for Phases 1-4 datasets (with outcomes between 2001 and 2013, minimum follow-up of 7 years).

Database	Fields
General hospital discharge record (SMR01) linked to mortality	Age in years Sex Admission date Admission type Admission reason Duration of hospital admission Days waiting ICD diagnostic group Main condition Other condition OPCS4 codes Main operation Date of the main operation Other operation Date of the other operation Date of death if dead Date of discharge Discharge type Discharge/transfer to Inpatient/day case marker Summarised admission code Summarised discharge code
Maternity and birth record (SMR02) linked to mortality	Age at admission (years) Age at conception (years) Type of antenatal care Total previous: Pregnancies Spontaneous abortions Therapeutic abortions Caesarean sections Still births Neonatal deaths Previous admissions this pregnancy Parity Date of booking Original booking Delivery plan-place Delivery plan-management Height of mother Mother height group Weight of mother Type of abortion Management of abortion Sterilisation after abortion Principal complication after abortion Estimated gestation Duration of pregnancy

Contd.

Database	Fields
	Antenatal steroids Diabetes Smoking history (booking) Smoker during pregnancy Condition on discharge Drug misuse (this pregnancy) Drugs used 1-4 Ever injected illicit drugs Typical weekly alcohol consumption Indication of labour Duration of labour Analgesia during labour/delivery
Psychiatric inpatient records (SMR04) linked to mortality	**As for SMR01 plus:** Status on admission Admission-referral from Previous psychiatric care Type of psychiatric care provided Age on discharge (years) Discharge-main condition Discharge-other condition 1-5 ECT 1st treatment date ECT treatments-no. this episode Clinical facility start Clinical facility end Arrangements for aftercare 1-4 Care plan arrangements
Cancer registry (SMR06) linked to mortality	Date of incidence/incidence date Side Site ICD9 ICD10 cancer site ICDO2 Type ICDO Morphology Date of death Vital status Embarkation date Cause of death 1-4 Death certificate only Grade classification Grade cell type MVB diagnosis Histological verification Method 1st detection Stage clinical T Stage clinical N Stage clinical M Stage colorectal Tumour size Nodes examined No of nodes examined Positive nodes Positive nodes no ER status Surgery Date 1st surgery Hosp GP 1st surgery Referred to radiotherapy Treated with radiotherapy Date of 1st radiotherapy Hospital 1st radiotherapy Chemotherapy Date 1st chemotherapy Hospital GP 1st chemotherapy

Contd.

Database	Fields
	Hormone therapy Date 1st hormone therapy Hosp GP 1st hormone therapy Other therapy Type other therapy Date 1st other therapy Hosp GP 1st other therapy
NRS death records	Date of event Primary cause of death Secondary causes of death
Scottish Breast Screening data	Date of death Cancer found Episode Episode type Episode invited Episode attended Episode outcome Cancer diagnosis date Facility flag Referral flag

Child health data	Birth details
	Birthweight
	Head circumference
	Number born (current pregnancy)
	Number born alive (current pregnancy)
	Birth order (current)
	Birth place type
	Age of mother
	Gestational age
	Delivery mode
	Onset of labour
	Location
	Early life/new born:
	  - Feeding at 6 weeks
	  - Neonatal care level
	  - Newborn screening-all
	  - Newborn status exam
	Health makers:
	  - Height
	  - Weight
	  - BMI
	  - Immusation activity
	  - History of present illness (?)
	  - Significant health concerns
	  - Sympathetic nervous system(??) status
	  - Dental status
	Basic health measurements
	  - Height
	  - Weight
	  - BMI

Primary care risk factor data	Date of registration Date of deregistration Tobacco consumption Tobacco consumption date Family history of disease Family history of disease date Exercise status Exercise date Diabetes Diabetes date CHD CHD date

Contd.

Database	Fields
	Stroke Strokes date Atrial fibrillation AF date Statins Statins date Height Height date Weight Weight date Cholesterol Cholesterol date Systolic blood pressure Systolic blood pressure date Diastolic blood pressure Diastolic blood pressure date Asthma Asthma date Asthma prescription Asthma prescription date
Scottish Bowel Cancer Screening Program linked to cancer registry (SMR06)	Date of test kit sent to participant Health board of residence Sex Age in years Screening test result Flag for kit completed in error Health board identifier/code Date of notification of a screening result Colonoscopy performed Date colonoscopy performed Reason for not having a colonoscopy Colonoscopy completed Invasive cancer detected ICD-10 classification of neoplasm Tumour classification (after surgery) Nodal classification (after surgery) Metastases classification (after surgery) TNM classification of malignant tumour (tumour/nodal status/metastasis) derived Dukes' stage Polyps detected Adenoma detected Count of adenomas Maximum dimension of the largest adenoma Polyp cancer detected Polypectomy performed at colonoscopy Complication from the colonoscopy requiring admission Death Site ICD9 ICD10 cancer site ICDO2 Type ICD 03 Grade cell type Stage colorectal
General hospital discharge record (SMR01) linked to hepatitis B dataset from Health Protection Scotland (HPS) and mortality	Sex Age at diagnosis (years) NHS board of residence (at diagnosis) Date of earliest HBs Ag positive specimen Source of 1st positive test (hospital, routine/antenatal screen, GP, other community setting) HBV test result (recent/acute, chronic) Date of late diagnosis Late diagnosis indicator Date of death

Contd.

Database	Fields
	As for Hepatitis B plus: Date of earliest positive specimen
General hospital discharge record (SMR01) linked to hepatitis C dataset from HPS and mortality	Source of 1st positive HCV test (hospital, routine/antenatal screen, GP, other community setting) HCV test result Description of result HCV genotype Risk group Date of first attendance (specialist services) Time from diagnosis to 1st attendance (specialist services) Date started antiviral therapy Time from diagnosis to start of antiviral therapy Response to antiviral therapy (sustained viral response(SVR) non-SVR)
General hospital discharge record (SMR01) linked to HIV dataset from HPS and mortality	**As for Hepatitis B plus:** Date of earliest positive specimen Source of HIV diagnosis test (hospital, routine/antenatal screen, GP, other community setting) Risk group Date of AIDS diagnosis (symptoms) New case (known in Scotland or elsewhere/new to Scotland/unknown) Infected outside to Scotland (yes/no) Follow up status (attending/not attending/lost/dead/left Scotland/recent) Date last attended healthcare services Date of 1st attendance in HIV specialist care Time from diagnosis to 1st attendance in HIV specialist care Date of 1st attendance for CD4 measurement 1st CD4 results (category: low, medium, high)

Table 3: Contents of Health Related Data.

Author contributions

SHELS coordinator: AD. SHELS principal investigator: RB. SHELS analyst: GC. Lead writer: JF. Co-writers: AD, CG, RB.

References

[1] **Bhopal R, Fischbacher C, Povey C, Chalmers J, Mueller G, Steiner M,** et al. 2010 Cohort profile: Scottish Health and Ethnicity Linkage Study of 4.65 million people exploring ethnic variations in disease in Scotland. Int J Epidemiol DOI: http://dx.doi.org/10.1093/ije/dyq118

[2] **Fischbacher CM, Bhopal R, Povey C, Steiner M, Chalmers J, Mueller G,** et al. 2007 Record linked retrospective cohort study of 4.6 million people exploring ethnic variations in disease: myocardial infarction in South Asians. BMC Public Health 2007; 7: 142. DOI: http://dx.doi.org/10.1186/1471-2458-7-142

[3] **Bansal N, Fischbacher CM, Bhopal RS,** et al. 2013 Myocardial infaction incidence and survival by ethnic group: Scottish Health and Ethnicity Linkage restropective cohort study. BMJ Open 2013; 3: e003415. DOI: http://dx.doi.org/10.1136/bmjopen-2013-003415

[4] **Bhopal RS, Bansal N, Fischbacher C,** et al. 2011 Ethnic variations in chest pain and angina in men and women: Scottish Ethnicity and Health Linkage Study of 4.65 million people. Eur J Cardiovasc Prev Rehabil 2011. DOI: http://dx.doi.org/10.1177/1741826711425775

[5] **Bhopal RS, Bansal N, Fischbacher CM,** et al. 2013 Ethnic variations in heart failure: Scottish Health and Ethnicity Linkage Study (SHELS). Heart 2012;98:468–73. DOI: http://dx.doi.org/10.1136/heartjnl-2011-301191

[6] **Bhopal RS, Bansal N, Fischbacher CM,** et al. Ethnic variations in the incidence and mortality of stroke in the Scottish Health and Ethnicity Linkage Study of 4.65 million people. 2011 Eur J Cardiovasc Prev Rehabil. Published Online First: 20 September 2011. DOI: http://dx.doi.org/10.1177/1741826711423217

[7] **Bhopal RS, Bansal N, Steiner M, Brewster DH,** 2012 "Does the Scottish Effect Apply to All Ethnic Groups? All-cancer, Lung, Colorectal, Breast and Prostate Cancer in the Scottish Health and Ethnicity Linkage Cohort Study." British Medical Journal Open 2 (5). pii: e001957. DOI: http://dx.doi.org/10.1136/bmjopen-2012-001957.

[8] **Sharpe KH, Cezard G, Bansal N, Bhopal R,** et al. 2013 Policy for home or hospice as the preferred place of death from cancer: Scottish Health and Ethnicity Linkage Study population cohort shows challenges across all ethnic groups in Scotland BMJ Supportive & Palliative Care 2013; 0: 1–9. DOI: http://dx.doi.org/10.1136/bmjspcare-2013-000485

[9] **Bansal N, Bhopal RS, Steiner MF,** et al. 2012 Major ethnic group differences in breast cancer screening uptake in Scotland are not extinguished by adjustment for indices of geographical residence, area deprivation, long-term illness and education. Br J Cancer 2012;106:1361–6. DOI: http://dx.doi.org/10.1038/bjc.2012.83

[10] **Bansal N, Chalmers J, Fischbacher, Steiner M, Bhopal R,** 2014 Ethnicity and first birth: age, smoking, delivery, gestation, weight and feeding: Scottish health and ethnicity linkage study. Eur J Public Health First published online: May 19, 2014 DOI: http://dx.doi.org/10.1093/eurpub/cku059

[11] **Bansal N, Bhopal R, Netto G, Lyons D, Steiner MFC, Sashidharan SP,** 2013 Disparate patterns of hospitalisation reflect unmet needs and persistent ethnic inequalities in mental health care: the Scottish health and ethnicity linkage study. Ethn Health 2013; 11: 1–23.

[12] **Fischbacher CM, Cezard G, Bhopal RS, Pearce J, Bansal N,** et al. 2013 Measures of socioeconomic position are not consistently associated with ethnic differences in cardiovascular disease in Scotland: methods from the Scottish Health and Ethnicity Linkage Study (SHELS). International Journal of Epidemiology 2013; 1-11 DOI: http://dx.doi.org/10.1093/ije/dyt237