Documentation for the archiving of the IID study data

Outline

1 Description of the originating project

- 1.1 Background to the study
- 1.2 Organisations involved in the study
- 1.3 Aims of the study

2 Study methods

Chapter 3 of the study report with full description of study methods

3 Structure and organisation of databases

- 3.1 Summary of stages of development of databases
  - 3.1.1 Study administration form databases
  - 3.1.2 Data collection management databases
  - 3.1.3 Questionnaire data databases
  - 3.1.4 Microbiology databases
  - 3.1.5 General practice information databases

4 Information on databases

- 4.1 Mapping of data collection, data entry and the creation of analysis databases, by study
- 4.2 Descriptive information on databases
- 4.2 Information on size of databases

5 Variable listings on all databases, with descriptions of all variables

Appendix containing coding listings of all 'open coded' questions for all questionnaires

6 Interpretation and coding of variables

- 6.1 General points applying to all datasets
- 6.2 Notes on coding of administration form data
- 6.3 Notes on coding of questionnaire data
  - 6.3.1 General points
  - 6.3.2 Coding notes relating to individual questionnaires
    Notes on additional variables for each questionnaire dataset
- 6.4 Notes on microbiology datasets

1 Description of the originating project

1.1 Background to study

In 1989, in response to national epidemics of foodborne infection with *Salmonella enteritidis* phage type 4, and *Listeria monocytogenes* the Secretary of State for Health and Minister of Agriculture, Fisheries and Food set up the Committee on the Microbiological Safety of Food, under the chairmanship of Professor Mark Richmond.[1-4] This committee recommended that

'a study of the incidence of infectious intestinal disease based on GP consultations in which microbiological confirmation of the clinical diagnosis is carried out'

and

'*the true incidence of infectious intestinal disease in the community needs to be ascertained. Thus we also recommend that a study including microbiological screening should be set up to provide information on the incidence of gastro-intestinal illness in the community that can be linked to a microbiological cause. This should take place, if possible in the same areas as the GP based study'*

In addition to these recommendations, the successors to the Richmond Committee decided that the value of the study would be enhanced by the collection of information on people without infectious intestinal disease, so that differences between the ill and the well could be identified. It was also decided that the clinical course of the disease, its long term sequelae and socio-economic costs should be addressed.

A pilot study was carried out in 1991-1992 which tested the feasibility of the design, established the basis for the sample size calculations, and compared options for selection and follow-up of subjects in general practice.[5]

Data collection for the IID study was started in September 1993 and completed in January 1996.

1.2 Organisations involved in the study[6]

- The Public Health Laboratory Service (PHLS) including the Communicable Disease Surveillance Centre (CDSC), London, Leeds Public Health Laboratory (PHL), and reference laboratories for specific organisms. These are: Laboratory of Enteric Pathogens (LEP) and Food Hygiene Laboratory (FHL), Central Public Health Laboratory (CPHL), London and the PHLS Anaerobe Reference Unit (Cardiff, PHL).
- The Centre for Applied Microbiology and Research (CAMR), Porton.

- The Medical Research Council (MRC) Epidemiology and Medical Care Unit (EMCU) and the MRC's General Practice Research Framework (GPRF).
- The Communicable Disease Epidemiology Unit and Health Services Research Unit of the London School of Hygiene and Tropical Medicine (LSHTM).

The organisations shared responsibility for the study: CDSC, EMCU and LSHTM were responsible for the design of the study; EMCU for the local organisation of the study in the general practices and the collection and entry of the data; LSHTM for the entry and analysis of data; Leeds PHL and LEP for the first line microbiological testing. Isolates were sent on to the relevant reference laboratories for confirmation and typing. CAMR was responsible for archiving isolates and stool specimens.

A Study Team consisting of representatives from EMCU, LSHTM and CDSC co-ordinated practice recruitment, nurse training, data collection within the practices, quality assurance, data processing and coding. Representatives from each of the laboratories met with microbiologists from the DH on a regular basis to review microbiological aspects of the study. Both groups reported to an Executive Committee. This met every three months to monitor progress and to advise on strategic and scientific issues.

1.3 Aims of the study[7]

The aim of the study was to estimate the number of cases of gastro-enteritis, or infectious intestinal diseases (IID), occurring in the population of England, and find out how many people with IID consulted their GPs and how these numbers compared with the numbers in national laboratory surveillance.

The study sought to identify as many as possible of the disease-causing organisms, or pathogens, responsible for IID. The estimate of the actual number of cases of IID in the population of England and presenting to their GPs, and the pathogens responsible for illness were compared with the routine national surveillance data from laboratory reports to the Public Health Laboratory Service (PHLS) Communicable Disease Surveillance Centre (CDSC). The study also set out to identify the factors which might lead to IID and the costs which might result.

Because it is impossible to separate out with any precision those cases of IID which result from food poisoning and those cases resulting from other causes, the study necessarily addressed **all** cases of IID and not merely the cases caused by eating contaminated food. Therefore, included in the study were cases infected with pathogens known to be spread predominantly from person to person, and pathogens usually held responsible for food poisoning, as well as those cases who, although clinically suffering from IID, had no pathogen found in their stools.

The study did not attempt to estimate the accuracy of national food poisoning statistics, which depend upon statutory notifications by doctors on the basis of clinical suspicion, but only of laboratory reporting to the PHLS CDSC.

The specific objectives of the study were:

- To estimate the number and aetiology of cases of IID in the population, presenting to GPs, and having stool specimens sent routinely for laboratory examination.

- To compare these numbers and the aetiologies with those recorded by the national laboratory reporting surveillance system.

- To estimate the prevalence of asymptomatic infection with agents associated with IID.

- To document differences between cases of IID (in the population and presenting to GPs) and similar people who were well (controls).

- To estimate the socio-economic burden of IID and its distribution.

References

1. Editorial: *Salmonella enteritidis* phage type 4: Chicken and egg. Lancet 1988; 2: 720-22.

2. Hopper TJ, Mead GC, Rowe B. Poultry meat as a source of human salmonellosis in England and Wales. Epidemiol Infect 1988; 100: 175-84.
3. Skirrow MB. Campylobacter enteritis: a new disease. BMJ 1977; 2: 9-11.
4. The Committee on the Microbiological Safety of Food. The microbiological safety of food: Part 1. London: HMSO, 1990.
5. Roderick PJ, Wheeler JG, Cowden JM, Sockett PN, Rodrigues LC. A pilot study of infectious intestinal disease (IID) in England. Epidemiol Infect 1995; 114:277-88.
6. Infectious Intestinal Disease Study Executive Committee. A Report of the Study of Infectious Intestinal Disease in England. 2000. London. The Stationery Office. Chapter 3 – Methods
7. Infectious Intestinal Disease Study Executive Committee. A Report of the Study of Infectious Intestinal Disease in England. 2000. London. The Stationery Office. Chapter 1 – Executive summary

2 Study Methods[6]

Study methods are discussed in detail in Appendix 1, the methods chapter from the final report of the IID study.

section 1	Study design
section 2	Stool collection and microbiology
section 3	Socio-economic component
section 4	Data-management and analysis
section 5	Monitoring performance

3 Structure and organisation of databases

There are 5 groups of databases, each associated with a different aspect of the study:

- Study Administration forms
- Data collection management databases
- Questionnaire data
- Data on microbiology
- Information on General Practices and age/sex distribution of practice lists

3.1 Generalised summary of stages of development from 'raw' datasets to datasets for analysis.

At the EMCU centre databases were developed using EPI INFO version 5.0 (Centers for Disease Control and Prevention (CDC), USA and World Health Organisation, Geneva, Switzerland). DBMS/COPY for Windows version 5.10.2 (Conceptual Software Inc., Houston, TX 77096) was used to transform data from one format to another. Analysis was carried out at LSHTM using Stata statistical software. Release 5.0. College Station, TX: Stata Corporation. 1997.

3.1.1 Study administration form databases

1. Data from each form was **double entered**, in EPI INFO, in batches of 50 forms. The data was validated and corrections made to one of the databases.

2. **Validations and preliminary checks** were made against administrative logging files at EMCU and with reference to practice nurses.

3. EPI INFO format files were **transformed** to STATA format, using DBMS/COPY.

3.1.2 Data collection management databases

There are 5 data collection management databases (EPI INFO files). There are separate databases for cases and controls, registering recruitment and there is a database for case recruitment to the enumeration study. The other two databases are log files of all questionnaires received and all stool specimens sent. These files were created and added to all through the process of data collection. As the different databases were checked some errors in these files were noted, checked and necessary alterations made.

3.1.3 Questionnaire data

1. Questionnaire data was **double entered**, in EPI INFO, in batches of 20 questionnaires, each batch to a separate database. The data was validated and corrections made to one set of databases. The corrected files were then merged using merge option 1 in EPI INFO, creating temporary files at each stage which cumulated to create the raw datasets.

2. **Validations and preliminary checks** were made against administrative logging files at the MRC and necessary changes made e.g. data entry errors corrected, blank records deleted. Programs run to identify inconsistencies in the data.

3. There were separate questionnaires for adults and children, cases and controls. For analysis, **datasets were merged** by appending in EPI INFO, to give two datasets, one with data on adult cases and controls and the other with data on child cases and controls.

4. EPI INFO format files were **transformed** to STATA format, using DBMS/COPY. There may be some variables renamed whilst transforming data files. If variable names in the EPI INFO file have more than 8 digits or if the first 8 digits of two variable names in the EPI INFO file are the same, the variable names in the STATA file will be renamed automatically.

5. Programs were run to implement consistency checks, making alterations to files.

6. ***A priori* recoding**. Programs were run to carry out some *a priori* recoding: all 'open code' listings were rationalised and coding reductions were carried out; new variables were created to make it possible to assess the risk of one factor which appeared in combination with other factors in the original coding e.g. a binary variable for ownership of a dog, regardless of ownership of other pets.

7. **Frequency based recoding**. Programs were run to recode variables based on descriptive summary of variables e.g. regrouping of categories where there were very low numbers in some categories.

3.1.4 Microbiology data

1. **Databases received from individual laboratories**. Preliminary checks and alterations made e.g. blank records and records of 'ineligible' specimens deleted.

2. Data received in files of different format: ASCII text, Excel (Microsoft Excel, Microsoft Corporation), DataEase, EPI INFO. Data has been **transformed** to STATA format for analysis.

3. **Analysis of individual datasets,** with creation of new variables.

4. **Files were merged** to give a final microbiology dataset containing the main results fields for all organisms. Some variables were dropped, new variables from encoded variables added, other new variables created and some recoding. All variable changes described in information on individual datasets

3.1.5 General practice information databases

Age and sex distributions of patients in each practice were collected at the start of each study cohort (2 per practice).

Information on the characteristics of each practice was also collected. Information on under-ascertainment and list-inflation factors was added to this database on completion of the appropriate analysis.

4 Information on databases

4.1 Mapping of data collection, data entry and creation of analysis databases, by study

Figure 1: Data map for the GP case control component

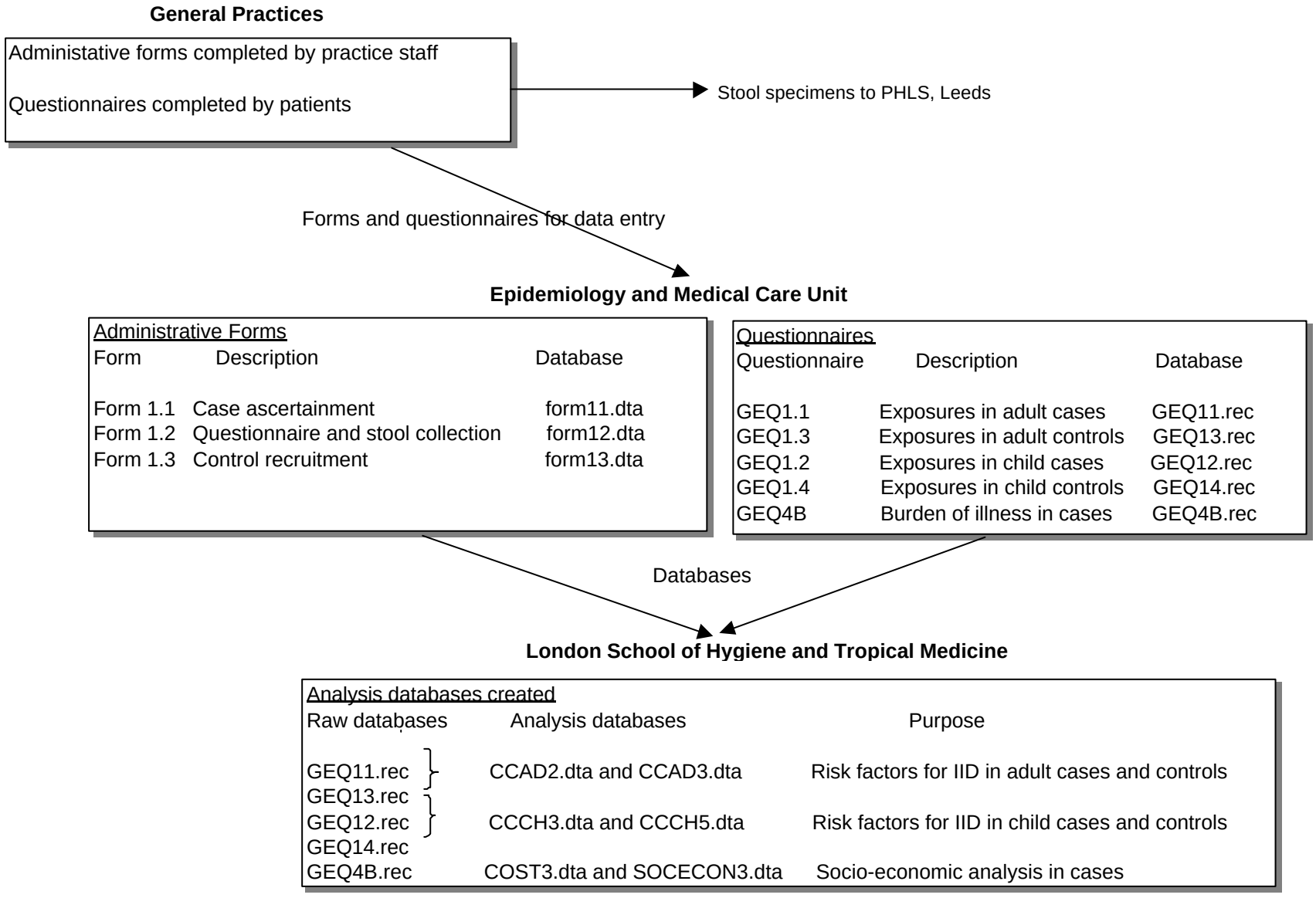


Figure 2: Data map for the Population cohort case control component

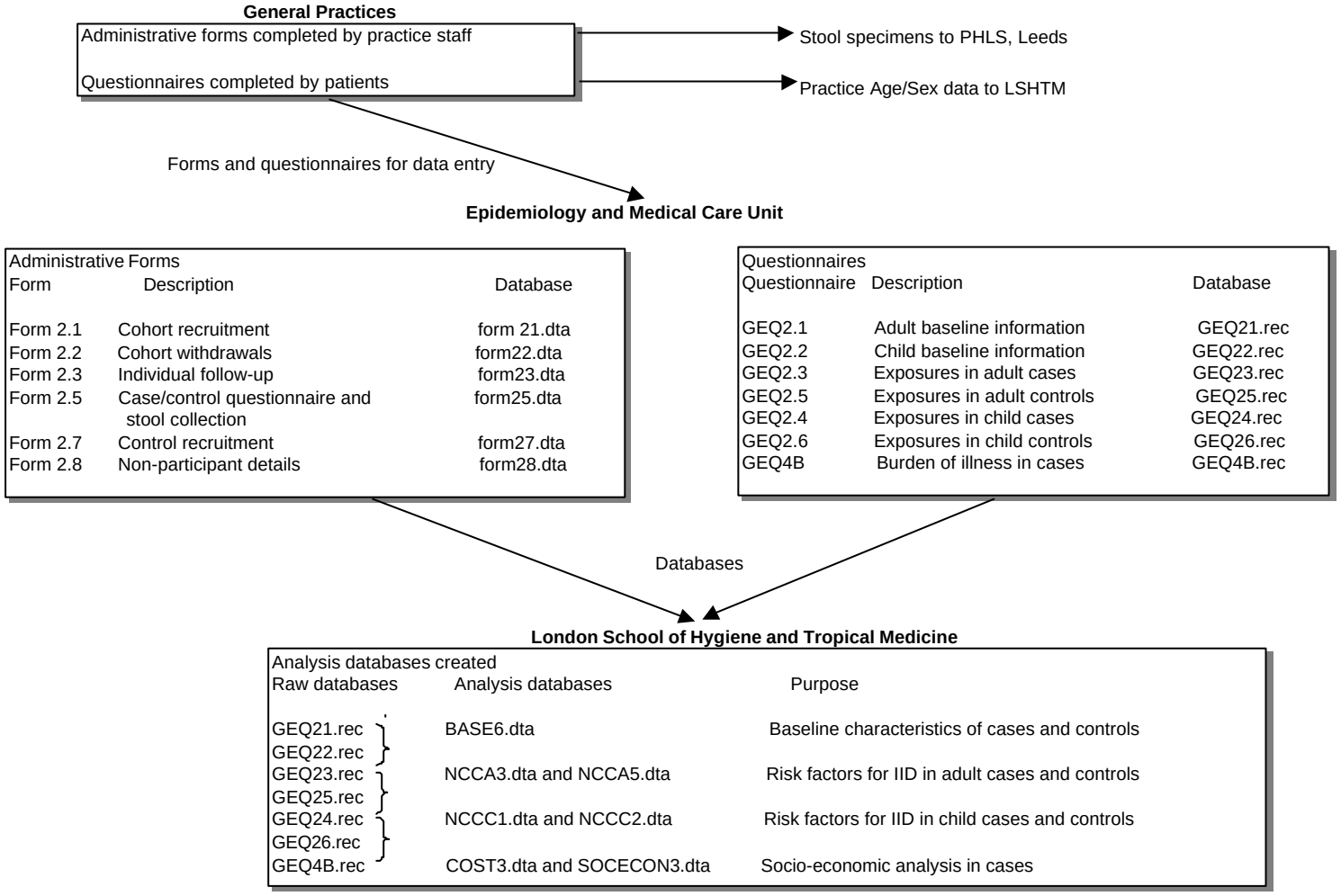


Figure 3: Enumeration component

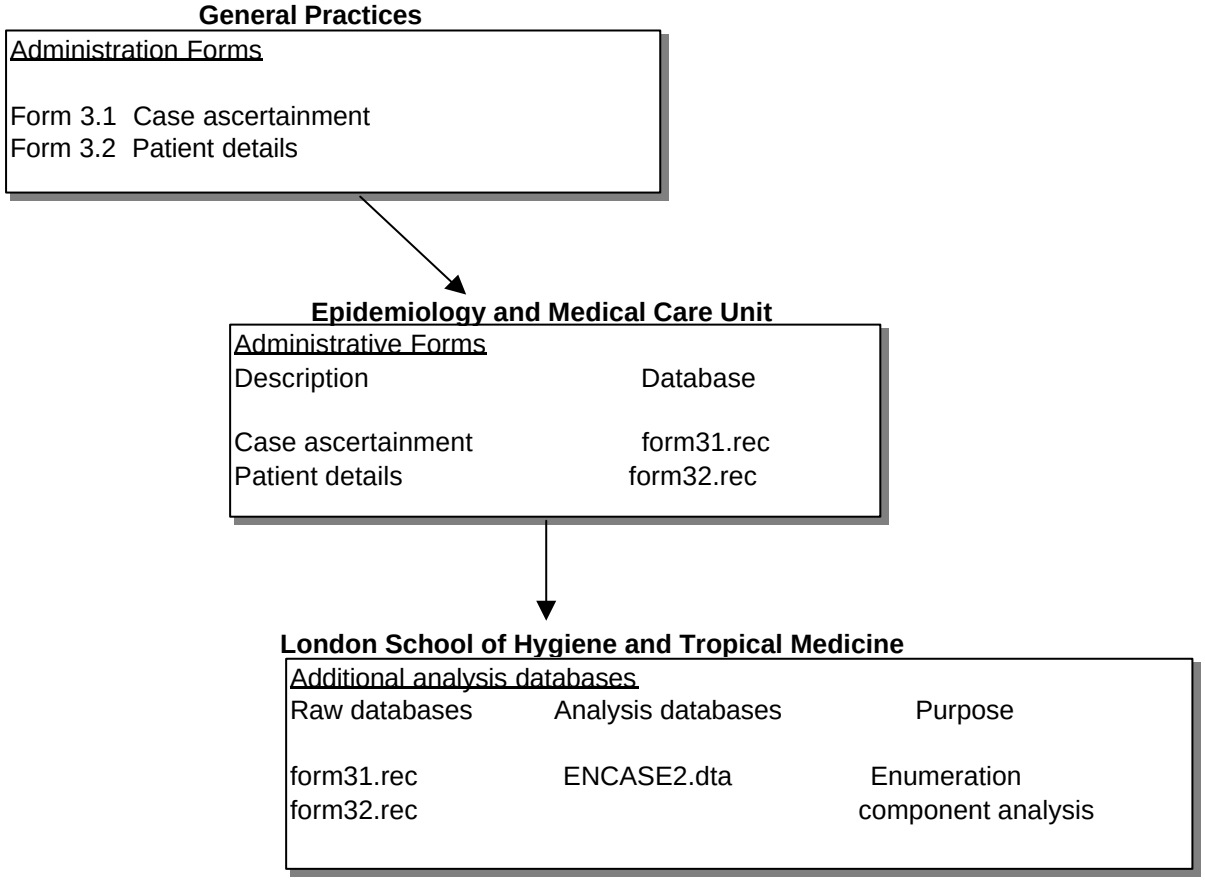
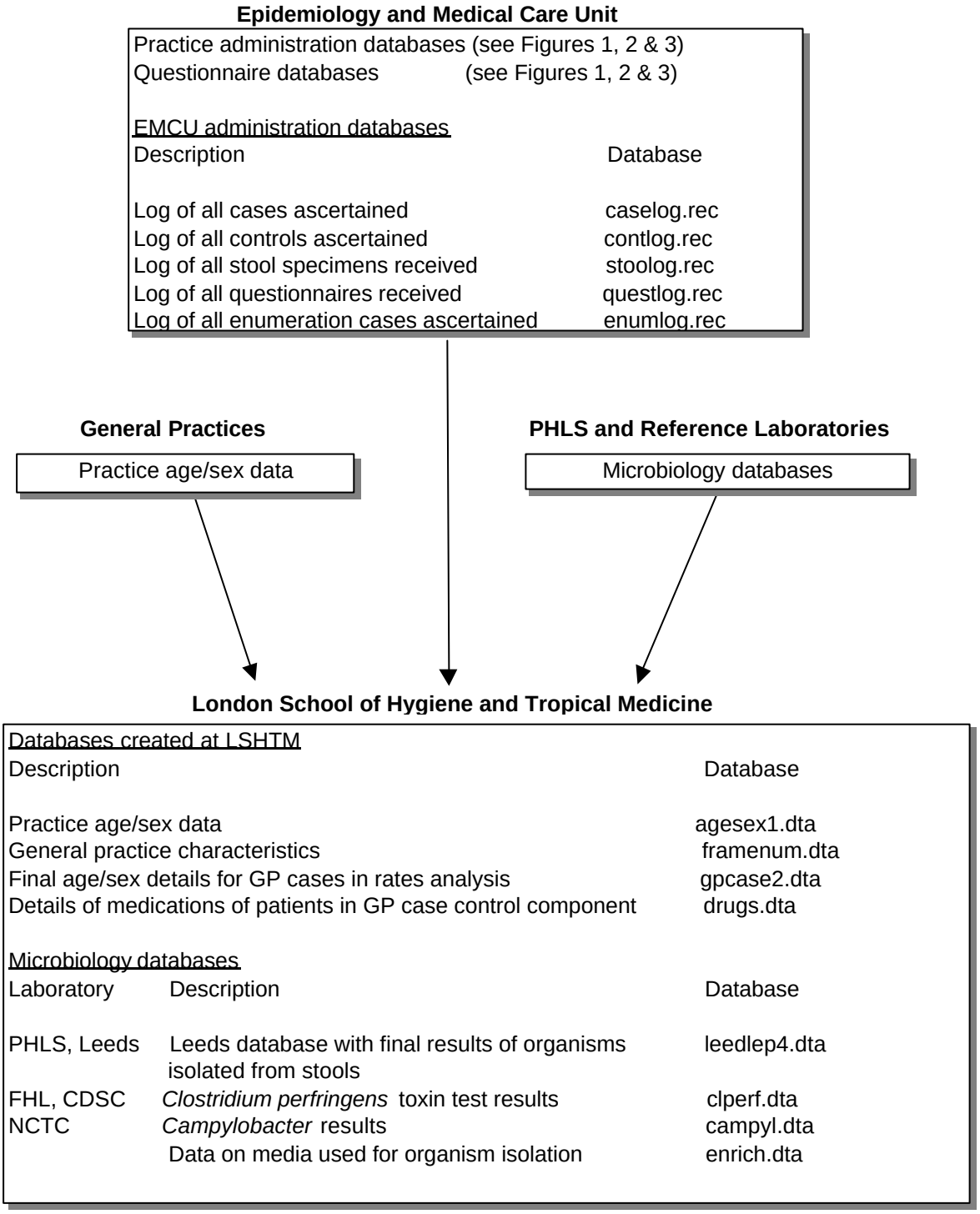


Figure 4: Map of routes of information sent to LSHTM and further databases created at LSHTM



4.2 Descriptive information on archived databases

IID Study Administrative Forms

Form number	Study component	Purpose	Datebase	'Final' database
1.1	Case control	Case ascertainment by GPs	FORM11. DTA	
1.2	Case control	Administer questionnaire & stool collection	FORM12. DTA	
1.3	Case control	Administer control recruitment	FORM13.DTA	
2.1	Cohort	Administer cohort recruitment	FORM21.DTA	
2.2	Cohort	Record cohort withdrawals	FORM22. DTA	
2.3	Cohort	Record cohort follow-up in individuals	FORM23. DTA	
2.5	Cohort	Administer case/control questionnaire & stool collection	FORM25.DTA	
2.7	Cohort	Administer control recruitment	FORM27.REC	
2.8	Cohort	Non-participant details	FORM28.DTA	
3.1	Enumeration	Case ascertainment by GPs	FORM31.DTA	ENCASE2.DTA - FORM31.DTA & FORM32.DTA merged
3.2	Enumeration	Patient details	FORM32.DTA	

IID Study Administration Databases

Study component	Purpose	Database
Case control & cohort	Log of all cases ascertained	CASELOG.REC
Case control & cohort	Log of all controls recruited	CONTLOG.REC
Enumeration	Log of all cases ascertained in the enumeration study	ENUMLOG.REC
Case control & cohort	Log of all stool specimens received	STOOLOG.REC
Case control & cohort	Log of all questionnaires received	QUESTLOG.REC

IID Study Questionnaires

SAQ number	Study component	Purpose	Analysis databases
1.1	Case control	Exposures in adult cases presenting to GP	CCAD2.DTA CCAD3.DTA
1.3	Case control	Exposures in adult controls presenting to GP	
1.2	Case control	Exposures in child cases presenting to GP	CCCH3.DTA CCCH5.DTA
1.4	Case control	Exposures in child controls presenting to GP	
2.1	Cohort	Adult baseline data	BASE6.DTA
2.2	Cohort	Child baseline data	
2.3	Cohort	Exposures in adult cases in the community	NCCA3.DTA NCCA4.DTA
2.5	Cohort	Exposures in adult controls in the community	
2.4	Cohort	Exposures in child cases in the community	NCCC1.DTA NCCC2.DTA
2.6	Cohort	Exposures in child controls in the community	
4B	Cost	Burden of disease in cases	COST3.DTA SOCECON3.DTA

IID Study Microbiology Databases

Laboratory	Description	Database
PHLS, Leeds	Leeds database with final results of organisms isolated from stools	LEEDLEP4.DTA
PHLS Leeds	Typing results on organisms tested at Leeds	LEEDTYPE.DTA
PHLS Leeds	*E.coli* typing results	LEPTYPE.DTA
FHL, CDSC Colindale	*Clostridium perfringens* toxin test results	CLPERF.DTA
NCTC	*Campylobacter* results	CAMPYL.DTA
PHLS, Leeds	Data on media used for organism isolation	ENRICH.DTA

Additional Databases

Description	Database
Age and sex data on all registered patients, by practice	AGESEX.DTA
General practice characteristics	FRAME.DTA
Final age/sex details for GP cases in the rates analysis	GPCASE2.DTA

4.3 Information on size of databases

Study administration form databases

File name	No. records	No. variables	Size of file (KB)
FORM11.DTA	3979	16	312
FORM12.DTA	3885	13	138
FORM13.DTA	3308	31	274
FORM21.DTA	27684	12	1083
FORM22.DTA	479	8	15
FORM23.DTA	9429	13	314
FORM25.DTA	756	13	34
FORM27.REC	707	24	75
FORM28.DTA	6529	6	128
FORM31.DTA	4541	12	152
FORM32.DTA	4750	15	196
ENCASE2.DTA	4876	26	342

Data collection management databases

File name	No. records	No. variables	Size of file (KB)
CASELOG.REC	4839	13	356
CONTLOG.REC	3208	11	208
ENUMLOG.REC	9249	6	290
STOOLOG.REC	6636	15	526
QUESLOG.REC	20760	10	995

Questionnaire databases

File name	No. records	No. variables	Size of file (KB)
CCAD2.DTA	2915	290	2666
CCAD3.DTA	2915	339	2948
CCCH3.DTA	2053	290	1884
CCCH5.DTA	2050	348	2117
BASE6.DTA	9466	85	1947
NCCA3.DTA	736	293	693
NCCA4.DTA	736	342	767
NCCC1.DTA	544	246	485
NCCC2.DTA	544	303	550
COST3.DTA	4529	359	6092
SOCECON3.DTA	4389	536	8703
DRUGS.DTA	3034	47	1452

Microbiology databases

File name	No. records	No. variables	Size of file (KB)
LEEDSLEP4.DTA	6473	78	1713
LEEDTYPE.DTA	6473	12	1210
LEPTYPE	774	18	78
CAMPYL.DTA	414	4	24
CLPERF.DTA	1557	118	1296
ENRICH.DTA	1062	27	216

Other databases

File name	No. records	No. variables	Size of file (KB)
AGESEX.DTA	70	77	53
GPCASE2.DTA	4026	10	182
FRAME.DTA	70	39	23

5 Variable listings with descriptions of all variables

See vol. 5 of documentation

6 Interpretation and coding of variables

6.1 General points applying to all datasets:

6.1.1 ID number and matching variables in the IID study

Each patient in the study has a unique ID number. The ID number is made up of indicators for study, practice, case/control status and individual study numbers. As well as the ID number, most datasets have variables for the separate components of the *ID* number i.e *study, practice, case* and *studynum*, and match variables i.e. *match1* and *match2*, which are substrings of the ID, to provide links between the datasets and between matched pairs within a dataset.

For every ID number the first digit indicates the study component:

1 - GP case-control component
2 - Population cohort component
3 - Enumeration component
This digit is reflected in the numbering of forms and questionnaires.

For every ID number the next 3 digits indicate the general practice research framework (GPRF) code [see list of practices, Appendix 3]

GP case-control component ID number e.g. 10240112

Digit	valid numbers	
1	1	study indicator
2 - 4	see list	GPRF code
5	0/1	control/case status
6 - 8	000 - 999	within each practice, the individual study number of the case within a matched pair. This number will be the same for the matched control.

In the GP case-control component *match1* variable is a 6 digit number, the ID number with the case/control status digit removed i.e. 1024112. This number will be the same for the case and control in a matched pair.

Population cohort component ID number e.g. 2031026

Digit	valid numbers	
1	2	study indicator
2 - 4	see list	GPRF code
5 - 7	000 - 999	within each practice, the individual study number

Population case-control component ID number e.g. 2031026011

The first 7 digits of this ID number are the ID number given to the patient at entry into the cohort.

Digit	valid numbers	
1	2	study indicator
2 - 4	see list	GPRF code
5 - 7	000 - 999	within each practice, the individual study number, allocated at baseline
8	0/1	control/case status
9 - 10	00 - 99	within each practice, the nested case-control component study number of the case within a matched pair, at entry into the cohort nested case-control component. This number will be the same for the matched control.

In the Population case-control component *match1* variable is a 7 digit number, the ID number with the case/control status digit and nested case-control component study number removed i.e. 2031026. This number is the ID number for the patient at entry into the cohort, so this number will link with the baseline questionnaire dataset.

In the Population case-control component *match2* variable is a 6 digit number, the study and practice numbers with the nested case-control component study number i.e. 203111. This number will be the same for the case and control in a matched pair in the nested case-control component.

Enumeration component ID number e.g. 3036058

Digit	valid numbers	
1	3	study indicator
2 - 4	see list	GPRF code
5 - 7	000 - 999	individual study number within practice

6.1.2 Comments/Anyqueries

Several databases have a comments or anyqueries field. The information in these fields was entered at data entry. The comments were used to highlight queries for data-cleaning. These fields have been left in the databases but the entries are not explained.

6.1.3 Anonymity of records

The datasets have been modified where necessary in order to prevent identification of general practices and subjects. The names of the towns of the general practices in the study and individual patient postcodes have been deleted.

6.2 Notes on coding of administration form data

Forms 1.1, 1.2
Codes for cause of non-infectious disease

Forms 1.2, 3.2
Codes for place of presentation to GP
D = deputising service, H = home visit, S = surgery, P = consultation by phone

Form 2.3
Form 2.3 contained an individual record of follow-up for 6 months. At data entry the data was collated and information on total number of weeks of follow-up and the week numbers for different events were entered onto the database.

Form 2.8
Coding for social class:

code	social class
1	I
2	II
3	III non-manual
4	IIImanual
5	IV
6	V
7	armed forces
8	not applicable
9	missing

Data from Forms 3.1 and 3.2 were merged for analysis. Data from the Form 3.2 was taken as correct where there were unresolved discrepancies between the data on the two forms, or if there was missing data on Form 3.1.

Information on sex, date of birth was checked for each ID. 8 records from practice 381 were dropped. This practice had a 'false start' when entering the study, so cases from this time period were not eligible for the study and their records were dropped.

Additional variables

f3231 - variable to indicate which forms were completed for each ID.
1 = Form 3.1 only, 2 = Form 3.2 only, 3 = Form 3.1 and Form 3.2
age and *agegroup* - age at consultation was calculated

6.3 Notes on coding of questionnaire data

The coding information relates to the three groups of questionnaires: GP case control component questionnaires (GEQ1.1, GEQ1.2,GEQ1.3, GEQ1.4), Community component questionnaires (GEQ2.1, GEQ2.2, baseline data and GEQ 2.3, GEQ2.4, GEQ2.5, GEQ2.6, nested case control component data) and socio-economic component questionnaire (GEQ4).

For each group of questionnaire databases there is a listing of variable names, types, label and a comment/description. In many cases the description on this listing should be sufficient to explain the coding of the variable on the database.

For some variables the comment refers to coding notes which give more specific details of how to interpret the coding for particular variables. These notes need to be read with reference to a particular questionnaire.

Some variables relate to 'open coded' questions i.e. question where the respondent wrote individual response rather than chose from options and the codes were developed during the study. The coding for these variables is found on the coding sheets relating to the particular questionnaire.

6.3.1 General points on coding:

Reserved codes
When coding the responses some numbers were reserved to represent a 'not sure/don't know' response, 'not applicable' or 'missing':

7, 77 or 777 were used as codes for 'not sure', when 'not sure' was given as a response to a question with an open code. (Other codes were sometimes used for 'not sure' when 'not sure' was given as an option in response to a question.)
8, 88 or 888 were used as codes for 'not applicable', where questions were nested e.g. in response to follow-on questions about pets for patients who had responded that they have no pets. (Similarly, other codes were sometimes used for 'not applicable' when 'not applicable' was given as an option in response to a question.)
9, 99 or 999 were used as codes for 'missing', were there was no response entered for a question.

Order of coding
Many questions have a set of options given and the respondent is required to tick one box e.g. GEQ1.1, Q.7.1 on marital status. There are 4 options, married, single, divorced or separated and widowed. The coding used is 1 - 4 with the codes given from left to right across the page, rather than in columns down the page e.g. married=1, single=2, divorced/separated=3 and widowed=4.

Social class coding
The information obtained from the 2 questions on occupation, job title and position at work e.g. Qs 7.3 and 7.4 on GEQ1.1, were used with reference to the OPCS Standard Occupational Classification (Office of Population Census and Surveys, Vols 1-3, HMSO 1993) to obtain a code for social class, based on occupation.

General coding rules

Where questions have several response options and more than one box was ticked by the respondent, when instructed to tick one only, e.g. GEQ1.1, Q. 9.8, the response was coded 0.

6.3.2 Coding notes relating to individual questionnaires

6.3.2.1 GEQ1.1, GEQ1.2, GEQ1.3, GEQ1.4

This information on coding is in addition to that given on the listing of variable names, types and descriptions for the questionnaire databases.

Q. 2.5, case symptoms
The response for each symptom is coded as a 4 digit number. The first 2 digits represent the total number of days with symptoms. The third digit represents the severity of symptoms (1 = mild, 2 = moderate, 3 = severe). The fourth digit is an indicator of whether or not the illness was still present at time of completing the questionnaire (1 = yes, 2 = no).

Q. 3.1, household members
The code for each household member is a 5 digit number, each code being made up of a series of 4 responses. The first 2 digits are for the age, the third digit represents the sex, (M=1, F=2). The 4th digit indicates whether the household member is a permanent member (1), or a visitor (2) and the 5th digit indicates whether the person was ill with diarrhoea and vomiting (yes = 1, no = 2 and not sure = 3).

Q. 3.4, 4.5, 5.1, 5.5, 6.3, 6.16, 6.20 (GEQ1.2), 6.22 (GEQ1.2), 7.1 (GEQ1.2), 7.2 (GEQ1.1), 7.2 (GEQ1.2), 7.7 (GEQ1.1), 8.1, 8.4, 9.6, 9.11, 9.16, 9.20
These questions have a set of options which are coded as described above and then an open coded section to add alternatives to the given options. The coding for these questions uses the numbers for the options given for the first codes and then numbers carrying on from there for the further alternative responses e.g. GEQ1.1, Q.3.4, codes are 1 - 6 for the options given and then 7 - 62 for alternatives entered as 'other' locations where contact occurred.

Q. 4.7, 4.8, 6.6, 6.7, 6.9, 6.10 (GEQ1.2), 9.18, 9.19
These questions have 2 digit codes as each code represents 2 options. The first digit of each code 1, 2 or 9 represents the option with 2 categories e.g. U.K./abroad in GEQ1.1, Q.4.7. The second digit represents the option with more categories e.g. swimming pool/sea/river/lake/other in GEQ1.1, Q.4.7. Thus the codes used for GEQ1.1, Q. 4.7 are 11 - 15, 21 - 25 for the options given and then 26 upwards for further open coded alternative options.

Q. 5.4, 6.1, 6.2, 6.14, 6.15, 9.3
These questions have each response entered as a separate variable.

Q. 6.4
The coding for this question is: never = 0, once = 1, more than once = 7, how many times? = exact number given or 6, if number given is 6 or greater

Q. 9.4
Some responses to this question were given with a temperature in Centigrade and some in Fahrenheit. For the range [15 - 20 degrees] it was unclear whether the response was Centgrade or Fahrenheit, so these responses were recoded to missing response.

Q. 9.8
This question has each option coded as a separate variable. The responses are coded: always = 1, sometimes = 2, never = 3, not sure = 4.

Q. 9.10
This question has each option coded as a separate variable. The responses are coded: top/middle shelf = 1, bottom shelf = 2, salad drawer = 3, in the door = 4, wherever there is room = 5, n/a = 6.

Q. 9.12
The response to this question should have been number of hours or ‘not applicable’. Some respondents entered ‘overnight’ on the questionnaire, so code (80) was introduced on 4/10/93 for ‘overnight’ thawing. Before this, overnight was coded as 12 hours. There were a few entries of 12 hours before this date which were recoded to missing as it was unclear what 12 hours meant (12 hours or overnight).

Q. 9.21
This question has each option coded as a separate variable. The responses are coded: agree = 1, disagree = 2, don’t know = 3.

GEQ4B

Q. 1.4, Characteristics of household members

The responses for each variable are coded separately with each characteristic entered as a separate variable. The sex of household members is not always known, if the sex response was missing and the relationship was for example 'friend' then the sex was not clear. In these cases the sex of the household member is coded for 'don't know'. The entry for occupation of household member is coded into categories related to social class (see coding list).

Q. 1.5

This question have each response entered as a separate variable.
There are 5 variables (OTH6W1 - 5) for up to 5 other symptoms.

Q. 1.6

This question has 5 variables (JTAFF1 - 5) for up to 5 joints listed as being painful/swollen.

Q. 2.8, 3.13, 4.5, 5.6, 6.6,

These questions have a set of options which are coded as described above and then an open coded section to add alternatives to the given options. The coding for these questions uses the numbers for the options given for the first codes and then numbers carrying on from there for the further alternative responses.

Q. 3.2, 4.4, 5.5, 6.4, 8.1, 9.6, Details of accompanying/caring person

The responses for each variable are coded separately with each response entered as a separate variable. The sex of the accompanying/caring person was taken from the response to the relationship of the accompanying/caring person. If this was for example 'friend' then the sex was not clear. In these cases the sex of the accompanying/caring person is coded as 'don't know'. The entry for occupation of household member is coded into categories related to social class (see coding list).

Q. 7.2
The other means of travel are coded 1 - 4 (see coding list). The further responses have separate variables for each entry.

Q. 12.7
For parts a) and c) of this question the relevant variables are 2 digit codes. The first digit corresponds to tick = 1, no tick = 2 for the first part of the question. The second digit is either 1 - 4 or missing = 9, if the first digit is 1, indicating a ticked box, or the second digit is 8, for not applicable.

OTHER SYMPTOMS

GEQ4/4B: Q1.5

01	Runny nose	31	Sensitive Teeth
02	Lack of concentration	32	Dry skin
03	Not sure	33	Stomach trouble
04	Swollen neck glands	34	Arthritis
05	Haemorrhoids	35	Chest Pains
06	Itching	36	Chesty cough
07	Continuous nausea without vomiting	37	Piles
08	Sore Throat	38	Lack of bladder control
09	Hoarseness	39	Burning sensation/Back
10	Weak	40	Fits
11	Sore Bottom	41	Burst blood vessels
12	Sore Mouth	42	Urine dark and strong
13	Feeling cold	43	Indigestion + Heartburn
14	Have difficulty in sleeping	44	Cold
15	Boils	45	Ear Infection/Earache
16	Sickness	46	Depression
17	Disturbed vision	47	Unable to sleep
18	Stools remain green	48	Hot/Cold sweats
19	Chicken Pox	49	Thirst
20	Itchy Eyes	50	Varicose Veins
21	Cramp in Feet	51	Bronchitis
22	Sore Ribs	52	Smelly Stools
23	Constipation	53	Nappy rash
24	Tonsillitis	54	Shiver
25	Pain in Groin	55	Irregular Heartburn
26	Everything hurts	56	Breathlessness
27	Feverish	57	Excessive belching
28	Mood Swings	58	Cystitis
29	Didn’t like light	59	Face swell
30	Restlessness	60	Labyrinthitis

DESCRIPTION: OTHER SYMPTOM

GEQ4: Q1.5

GEQ4B: Q1.5

Code	Description
61	Exhaustion
62	Peeling skin on hands
63	Tender skin
64	Anxiety
65	Sore tongue
66	Chest infection
67	Sensitive to certain foods
68	Abdominal swelling
69	Symptoms on & off
70	Pneumonia
71	Period started a week early
72	Mucus in stools
73	Sinusitis
74	Clammy hands
75	Developed phlebitis
76	Unable to eat
77	Coughing up phlem
78	Blocked nose
79	Passing water more frequently
80	Discomfort
81	Bad taste in mouth
82	Diabetes allergy
83	Painful bowel movements
84	Temporary deafness
85	Feeling sick
86	Stools light in
87	Loose stools
88	Bringing up blood when vomiting
89	Difficult to digest fatty foods
90	Acidity
91	Gingivitis (gum infection)
92	General listlessness
93	Conjunctivitis

RELATIONSHIP OF ACCOMPANYING PERSON

GEQ4: Q1.4, Q3.13, Q4.5, Q5.7, Q6.5, Q8.1
GEQ4B: Q1.4, Q3.12, Q4.1, Q5.5, Q6.4, Q8.1

Code	Description	Code	Description
1	Mother/Father	28	Niece/Nephew/Brother/Sister in law
2	Husband/Wife	29	Step Brother/Sister
3	Brother/Sister	30	Cousin
4	Son/Daughter	31	Brother/Sister in-law
5	Boyfriend/Girlfriend	32	Mother/Father partner
6	Grandparents	33	Foster parents
7	Friend	34	Nursing home/School staff
8	Mother/Father in-laws	35	Family
9	Grandchildren	36	Boyfriend/Girlfriend Parents
10	Son/Daughter In-laws	37	Great Grandparents
11	Neighbour	38	Housemaid
12	Aunt/Uncle	39	Employee
13	Step parents	40	Ex Husband/Wife
14	Common-law Husband/Wife	41	Friends Parents
15	Flatmate/Lodger	42	Doctor
16	Landlord (house)		
17	Babysitter / Au Pair		
18	None		
19	Son/Daughter Boyfriend/Girlfriend		
20	Carer		
21	Teacher		
22	Unknown		
23	Employer		
24	Son/Daughter's Boyfriend/Girlfriend		
25	Work college		
26	Step Son/Daughter		
27	Great Aunt/Uncle		

JOINTS AFFECTED

GEOQ4/4B: Q1.6

01	Legs		
02	Shoulder		
03	Neck		
04	Arms		
05	Wrists		
06	Ankles		
07	Knees		
08	Back		
09	Hips		
10	Elbow		
11	Hands		
12	Heals		
13	Feet		
14	All joints		
15	Toes		
16	Thighs		
17	Spine		
18	Pelvic		
19	Calves		
20	Finger		
21	Knuckles		

ABSENCE AFFECTED WORK

GEQ4/4B: Q2.8

05	Colleague/Temporary Staff		
06	Colleague/Until returned		
07	None of Above		
08	Employed on (Bank) basis		
09	Retired, On holiday, Unemployed, YT Training Maternity leave		
10	Have to copy/not miss work		
11	Self employed		
12	Got the sack		
13	Work had to be cancelled		
14	Colleague/until returned/temp.		
15	Self employed / contact others		
16	Work extra day		
17	Had to leave work		
18	Until returned/colleague/work at home		
19	Until returned/work at home		
20	Work at home		
21	Until returned/Temp. staff		
22	Work Part Time so had to adjust hours		
23	Had to take holiday		
24	I'm a Nanny so mother had to stay home		
25			
26			
27			
28			
29			
30			

ARRANGEMENTS OF ACCOMPANYING PERSON

GEQ4: Q3.14, Q4.6, Q5.8, Q6.7
GEQ4QB: Q3.13, Q4.5, Q5.6, Q6.6

01	None		
02	Time off Work		
03	Baby sitter or carer for someone		
04	Time off school/college		
05	Cancel planned arrangements		
06	Other not specified		
07	Not known		
08	Was already off work (holiday)		
09	Someone to drive		
10	Cancel child minder		
11	Time off work/cancelled arrangements		
12	Time off school/cancelled arrangements		
13	Babysitter/time off work		
14	Babysitter/cancelled arrangements		
15	Collect tablets from chemist		
16	The person was ill themselves		
17	Cancelled arrangements/arrange transport		
18	Had to make lost time up at work		
19	Someone to cover me at work		
20	Time off work/college/cancelled arrangements		
21	Time off work/cancelled arrangements		
22	Time off work/transport		
23	Stayed at home		
24	Someone to drive babysitter		

OTHER

GEQ4B: Q5.3

01	Flowers		
02	Taxi		
03	Travel		
04	T.V.		
05	Petrol		
06	Parking		
07	Baby wipes		
08	Nappies		
09	Yellow brick road		

OTHER MEANS OF TRAVEL

GEQ4/4B: Q7.2

01	Walking		
02	Cycle		
03	Motorcycle		
04	Electric scooter		

OTHER COSTS IN GENERAL

GEQ4/4B: Q10.13

001	Telephone calls	029	Cost to company car hire
002	None	030	Electric blanket
003	Calpole	031	Frustrations (being ill)
004	Sugar and salt	032	Social engagements
005	Petrol	033	Mattress
006	Loss of work/self employed	034	Consultancy fees
007	Food	035	Not able to do shopping
008	Drinks/orange squash	036	Prescription fees
009	Heating	037	Electricity
010	Taxi		
011	Baby minding fee		
012	Petrol, using private car		
013	Medicines		
014	Doctors sick note		
015	Nappy sacks		
016	Creams (medicated)		
017	Lost holidays		
018	Extra washing		
019	Stamps		
020	Travel costs		
021	Car parking bills		
022	Toilet paper		
023	Water meter bills		
024	Lost my job		
025	Missed exams		
026	Nappies		
027	Gas		
028	Recreation		

OTHER COSTS ON HOLIDAY

GEQ4B: Q9.14

01	Telephone calls		
02	None		
03	Medicine		
04	Wasted food		
05	Excess to pay		
06	Doctor call out charge and prescriptions		
07	Phone calls and taxi fares		
08	Laundry		
09	Doctors fees and blood test		

OCCUPATION CODES

GEQ4B:

01	I
02	II
03	III N
04	IV M
05	IV
06	V
07	VI UNEMPLOYED
08	VII other including forces/ mother/disabled
09	HOUSEWIFE
10	STUDENT (17 up)
11	CHILD (16 down)
12	RETIRED

Notes on additional variables

The data from the socio-economic questionnaire was merged with data from other datasets to obtain information necessary for analysis e.g. data on age, sex and social factors and information on organisms isolated. These additional variables are listed and described on the database information sheets. Several new variables were created before analysis, these are described below.

weeks - time interval between onset of illness and completion of questionnaire

Criterion for exclusion of cases from the socio-economic analysis

Questionnaires should have been given to patients 3 weeks after their illness.
Estimates of the time interval between illness and completion of questionnaires (*weeks* variable) showed a range of [1 - 35 weeks]. By comparing this time interval with the number of days patients reported being unwell 3 categories of cases whose data was inconsistent with the study protocol were identified.

a) Number of days unwell was greater than the time interval between consultation date and completion of questionnaire.
80% of cases in this category were from the Enumeration study. It was thought that patients in the Enumeration study were more likely to have been ill for a while before consulting a doctor. However, the study protocol excludes patients who have had loose stools for more than 14 days duration at presentation to the GP. 31 cases were excluded from the analysis who appeared not to fulfil the study criteria for inclusion.

b) Time interval between onset of illness and completion of questionnaire was less than 3 weeks.
The majority of the cases in this category were in the Enumeration study where it appears that the patients were given the questionnaire before 3 weeks and then completed it after they had recovered from their illness. In all but 1 case in this category the patient was unwell for a few days and completed the questionnaire after they had recovered. 1 case was excluded as the onset date and completion date were the same.

c) Patients returning their questionnaires late.
90% of patients had returned their questionnaires by 8 weeks, 95% by 10 weeks. It was decided to include cases who had completed their questionnaires late. Some appeared to have been unwell for a long period and completed their questionnaires on recovery, others had not been unwell for long but had been late in returning their questionnaires. It was noted that recall may make the data less reliable for these patients, but they were not excluded.

Some new variables were encoded versions of previous variables e.g. age and sex variables for household members. The encoding changes the type of variable from a string/character variable to a numeric variable that can be used in analysis.

hhage - indicator variable for the number of adults and children in a household

This variable has a 2-digit code, the first digit indicates the number of adults in the household and the second digit indicates the number of children in the household.

hhsex - indicator variable for the sex of the members of the household

This variable has a 2-digit code, the first digit indicates the number of females in the household and the second digit indicates the number of males in the household.

Both these variables give information on the members of the household excluding the case.

month and *year* - month and year of onset of illness used in the calculation of code for time category relating to prescription costs

presc2 - codes 1 - 3 relating to time interval for use of different prescription rates (see methods section for table of cost vectors)

The variable exempt was cleaned to include children under 16, women over 60 and men over 65 year of age as exempt from payment of prescription costs. The prescription costs of cases who were not exempt were considered as personal costs. The prescription costs of cases who were exempt were attributed to the NHS.

othanycost - information on additional other personal costs was recoded into 11 categories.

code	description
40	communications
41	medical expenses
42	food/food supplements
43	transport
44	work
45	fuel
46	care
47	hygiene
48	leisure
49	school
50	cost in the home

6.4 Notes on Microbiology databases

LEEDSLEP4.DTA – Leeds database

Coding for variables org1, org2, org3, org4

Abbreviation	Name of organism
ad	Adenovirus
ae	*Aeromonas*
as	Astrovirus
ba	*Bacillus*
ca	*Campylobacter*
cd	*Cl..difficile*
cr	*Cryptosporidium*
cv	Calicivirus
e1	Enterotoxigenic *E.coli* (ETEC)
e2	Verocytotoxigenic *E.coli* (VTEC, non O157)
e3	Enteroinvasive *E.coli*
e4	Attaching & effacing *E.coli* (AEEC)
e5	Enteropathogenic *E.coli* (EPEC)
e6	Enteroaggregative *E.coli* (EaggEC)
e7	Diffusely adherent *E.coli* (DAEC)
ec	*E.coli* O157
gi	*Giardia*
rc	Rotavirus Group C
rv	Rotavirus Group A
sa	*Salmonella*
sh	*Shigella*
sr	SRSV
st	Staph.aureus
vb	*Vibrio*
ys	*Yersinia*

CLPERF.DTA - *Clostridium perfringens*

This dataset contains the results data on *Cl.perfringens* both from the Leeds laboratory and from the Food Hygiene Laboratory, the reference laboratory for Clostridia. It also contains variables that have been created from the results data. These notes are in addition to the variable descriptions on the database information sheets.

The Leeds laboratory sent various types of specimen to the reference laboratory. This means that each patient id may have information on several specimens (up to 11 per patient id). The information on each specimen is contained in 10 variables:

etlds	RPLA result from Leeds equivocal/negative/other/positive
etfhl (etf)	enterotoxin result at FHL (Negative)/(Positive) - RPLA result Negative/Positive - ELISA result Insufficient
elisa	ELISA result at FHL FcI/Insufficient/negative/positive/end
etrpla (etrp)	RPLA result at FHL
ident	identification of other organisms isolated
sero	serotype of *Cl.perfringens*
morph	morphology of the isolate rough/smooth
source (sour)	source of the antigen for identification spore/viable

Each specimen has a set of these variable with a number (1 - 11) after the variable name to indicate which specimen.

Additional variables have been created to summarise the data e.g. specno, the number of specimens per ID number. In order to do this some variables have been encoded, changing the type from string/character variable to numeric e.g. the source of specimen variable and the etfhl and etrpla variables (see above for description). It is important to note that the coding for these additional variable does not always represent the same result, as the coding will reflect the listing for a particular variable e.g. etf4 = 1 is negative whereas etf5 = 1 is positive as there where no negative results for these samples.

clpfpos original Leeds result on the specimens sent to FHL except for those where a different organism was isolated at the FHL or there was insufficient specimen to test at the FHL.
0 = negative, 1 = positive, 2 = other 3 = equivocal

6.5 Link to the archive of stool specimens at CAMR

The database of the laboratory results (LEEDLEP4.dta) has variables ID (individual identification number) and LABNO (individual Leeds laboratory number) with which each stool specimen can be identified. Also on the database is the 'Yes/No' variable SENTCAMR (specimen sent to CAMR for archiving).

Further information will be available confirming the ID numbers with specimens in the archive at CAMR and the nature of the specimens available.

Publications (full listing from Andrea Belcher at FSA)

1. Roderick P, Wheeler J, Cowden J, Sockett P, Skinner R, Mortimer P, Rowe B, Rodriques L.
   A pilot study of infectious intestinal disease in England. Epidemiol Infect. 1995 Apr;114(2):277-88.

2. Sethi D, Wheeler J, Rodrigues LC, Fox S, Roderick P.
   Investigation of under-ascertainment in epidemiological studies based in general practice. Int J Epidemiol 1999; 28: 106-12.

3. Wheeler JG, Sethi D, Cowden JM, Wall PG, Rodrigues LC, Tompkins DS, Hudson MJ, Roderick PJ.
   Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. BMJ. 1999 Apr 17;318(7190):1046-50.

4. Tompkins DS, Hudson MJ, Smith HR, Eglin RP, Wheeler JG, Brett MM, Owen RJ, Brazier JS, Cumberland P, King V, Cook PE.
   A study of infectious intestinal disease in England: microbiological findings in cases and controls. Commun Dis Public Health. 1999 Jun;2(2):108-13.

5. Sethi D, Wheeler JG, Cowden JM, Rodrigues LC, Sockett PN, Roberts JA, Cumberland P, Tompkins DS, Wall PG, Hudson MJ, Roderick PJ.
   A study of infectious intestinal disease in England: plan and methods of data collection. Commun Dis Public Health. 1999 Jun;2(2):101-7.

6. Sethi D, Cumberland P, Hudson MJ, Rodrigues LC, Wheeler JG, Roberts JA, Tompkins DS, Cowden JM, Roderick PJ.
   A study of infectious intestinal disease in England: risk factors associated with group A rotavirus in children. Epidemiol Infect. 2001 Feb;126(1):63-70.

7. Rodrigues LC, Cowden JM, Wheeler JG, Sethi D, Wall PG, Cumberland P, Tompkins DS, Hudson MJ, Roberts JA, Roderick PJ.
   The study of infectious intestinal disease in England: risk factors for cases of infectious intestinal disease with Campylobacter jejuni infection. Epidemiol Infect. 2001 Oct;127(2):185-93.

8. Evans J, Wilson A, Willshaw GA, Cheasty T, Tompkins DS, Wheeler JG, Smith HR.
   Vero cytotoxin-producing Escherichia coli in a study of infectious intestinal disease in England. Clin Microbiol Infect 2002 Mar;8(3):183-6.

9. A. Wilson, J. Evans, H. Chart, T. Cheasty, J.G. Wheeler, D. Tompkins & H.R. Smith
   Characterisation of strains of enteroaggregative *Escherichia coli* isolated during the infectious intestinal disease study in England. European Journal of Epidemiology 2002 17(12): 1125-1130

IID study report – see other documentation

User queries: Contact Andrea Belcher at FSA