 HIV Medicine (2004), **5**, 115–124

ORIGINAL RESEARCH

The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study

The UK Collaborative HIV Cohort Steering Committee*

Objectives

This paper describes the development of the UK Collaborative HIV Cohort (CHIC) Study. The aim of the study is to collate routinely collected data on HIV-infected individuals attending one of seven clinical centres in the UK since 1 January 1996, with the objectives of describing changes over time in the frequency of AIDS-defining illnesses, describing the uptake of and response to highly active antiretroviral therapy (HAART), and identifying factors associated with virological and immunological responses to HAART.

Methods

By December 2002, demographic, clinical and laboratory data had been collected on HIV-positive patients seen at six of the seven HIV centres. Missing and inconsistent data had been investigated and the datasets audited. Records identified as relating to the same patient had been merged, and cross-checks made with UK death registers to improve the accuracy of death reporting.

Results

The cohort currently contains information on 13 833 individuals. Eighty-two per cent of the cohort are male, and the median age was 34 years at first follow-up. The main risk factors for HIV infection have been determined as sex between men (63%) and sex between men and women (24%). Twenty-five per cent of the cohort are known to have developed AIDS, and 8% have died.

Conclusions

The UK CHIC Study provides important information on the status of individuals infected with HIV in the UK, and provides a means to study the response to HAART and to monitor changes in the clinical event and death rates that have occurred since the introduction of HAART in the UK.

Keywords: cohort study, epidemiology, HIV, natural history, observational database

Introduction

From a public health perspective, it is important to continue to monitor the outcomes of individuals infected with HIV at a national level. Whilst existing surveillance systems in the UK collect information on significant data items, such as new HIV diagnoses, newly diagnosed AIDS cases and deaths [1], the amount of information collected by such systems is, by necessity, restricted and this limits the questions that can be answered. A combined cohort comprising data from a large number of diagnosed HIV-infected individuals in the UK would provide not only basic information on the status of individuals infected with HIV, but also data to address important questions relating to the natural history of disease and the impact of antiretroviral therapy.

Whilst many of the larger clinical centres in the UK already maintain computer databases that can be used to address such issues, the benefits of national HIV databases have long been recognized [2–4]. In particular, the outcomes of studies from a single centre and the generalization of the findings may be restricted by insufficient numbers of patients, by the demographic breakdown of individuals attending a centre, or by centre-specific protocols or treatment practices. A combined UK database would have the advantage of providing long-term follow-up data on a large and diverse cohort of patients, and would provide increased statistical power for studies of less

Correspondence: Dr Caroline Sabin, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, Rowland Hill St, London NW3 2PF, UK. Tel: +44 (0) 20 7830 2239; fax: +44 (0) 20 7794 1224; e-mail: c.sabin@pcps.ucl.ac.uk

*See Appendix 1.

commonly occurring clinical events. Also, linking of data on patients attending different centres reduces any bias that may arise from loss to follow-up.

The UK Collaborative HIV Cohort (UK CHIC) Study was initiated in 2001 following discussions between clinicians, epidemiologists and statisticians about the requirements of a large UK-based cohort. The overall aim of the collaboration is to create a large cohort of individuals receiving care for HIV in the UK, with specific objectives of describing changes over time in the frequency of AIDS-defining illnesses, describing the uptake of and response to highly active antiretroviral therapy (HAART), and identifying factors associated with virological and immunological response to HAART. The collaboration currently involves seven of the largest clinical centres that provide medical care to HIV-infected individuals in the UK, six of which are situated in London (Chelsea and Westminster NHS Trust, King's College Hospital, Mortimer Market Centre, St. Mary's Hospital, St. Thomas' Hospital and the Royal Free Hospital) and one in Brighton (Brighton and Sussex University Hospital). Of the individuals reported as seen for HIV-related care in the UK during 2001, approximately 40% are thought to have attended for care at one of these centres (Lara Payne, Health Protection Agency, personal communication). Many of the centres have already presented results on their own datasets [5–8] or have collaborated in combined cohort studies on a wide range of topics [9–12].

The aim of this paper was to describe the development of the cohort, the methodological approaches that have been taken and the demographic and clinical characteristics of the patients currently enrolled in the cohort.

Methods

Organization and ethics

The UK CHIC Steering Committee is made up of 13 clinicians, epidemiologists and statisticians representing the seven centres that agreed to contribute data, the MRC Clinical Trials Unit and the Health Protection Agency-Communicable Disease Surveillance Centre (HPA-CDSC) (Appendix 1). The project was approved by a Multi-centre Research Ethics Committee (MREC) and by local ethics committees.

Data collection

The criteria for inclusion of an individual in the UK CHIC Study were that a person was HIV positive, was aged over 16 years and had attended one of the centres for care at any time in 1996 or thereafter. This year was chosen as the introduction of HAART occurred in the UK around this time.

The data items to be collected, their format and the feasibility of obtaining these from clinic databases were discussed and agreed upon by representatives from all centres. Each centre provided electronically formatted data in specified datasets: demographic information; AIDS diagnoses; laboratory data (CD4 and CD8 counts, and viral loads) and antiretroviral treatment (ART) history (complete variable list available from the corresponding author on request). Data were provided only for those patients seen at a centre from 1 January 1996 onwards, although if an individual had attended before this time then data from the time of first attendance were collected, thus providing a complete clinical history whilst at that centre. Subsequent downloads to update and expand the database will occur every 6–9 months.

In accordance with data protection policy, data were provided by clinics in a pseudo-anonymized format with all names removed and replaced by first-name initial and a soundex code derived from the patient's surname [13]. Locally assigned clinic identification numbers provided the link to records in clinics, but these were removed and replaced by a unique study identifier after the data were merged and before the data were distributed for analysis.

Data queries/data improvement

In order to improve the accuracy and completeness of the data before any analyses were performed, an extensive range of queries was applied to the data from each centre. These queries identified missing or invalid demographic data items, particularly dates of birth and soundex codes, or data items that were illogical or that conflicted (for example, an HIV-positive test date occurring before an HIV-negative test date; a last-seen date occurring before the first-seen date; any dates for laboratory tests, AIDS events or start of ART occurring after a last-seen date or date of death). In order to improve the source dataset from each centre, the results of queries generated were reported back to the appropriate centre and were resolved where possible by examining the clinic database or patient records. Where more accurate information was established, both the source database and the UK CHIC database were updated, so that future data downloads would contain the verified data.

Linkage to death registries

As some individuals were lost to follow-up at centres, the Office for National Statistics (ONS) for England and Wales, and the General Registrar Office for Scotland (GRO) death

registers were used to ascertain whether or not these patients had subsequently died. Any UK CHIC records that matched a record in either the ONS or GRO database on first-name initial, soundex, date of birth and sex were identified. If this matching process identified subjects in the cohort recorded as having died, but with a missing date of death, the date of death from the ONS/GRO database was used in its place. Matches that occurred for individuals who were not recorded as having died were reported to the centres for verification before being updated on the database. As part of this process, information was obtained on 358 deaths that had not previously been known to the individual centres, and more accurate dates of death obtained for a further 141 individuals.

Data audit

To ascertain the degree to which data held in clinic databases accurately reflect what is written in clinical case notes, a random selection of 1% of patient records from each centre were audited. For each individual in the subsample, the dataset was recreated directly from patient notes by one investigator and was compared to the database record held for these individuals. Information on the outcome of the audit was then fed back to individual centres so that they could rectify any problems.

Date of birth, soundex, first-name initial, sex, exposure group, ethnicity and country of origin were expected to match exactly with the information found in patient notes. First-seen dates were assessed as being accurate if they matched to within 1 month of each other and HIV-positive and HIV-negative dates if they were within 1 year of each other. AIDS events were considered accurate if the diagnoses matched exactly and the dates of clinical events matched to within 1 month. AIDS events occurring in a patient's clinical history before their first-seen date at a centre were not audited. Individual drug records in an ART history were assessed as accurate if the drugs matched between the database and patients' notes and the start dates were within 1 month (if this date occurred after they were first seen at a centre).

Duplicate records

It is known that some individuals will have attended more than one centre for their care, resulting in two or more records relating to a single individual in the cohort. It was thought essential to combine information for those patients where, on the basis of soundex code, date of birth and other clinical information, there was a clear match between records. Initially, potential matches were identified where both date of birth and soundex matched. Other demographic variables were then used to formulate a series of rules for use in a computerized algorithm to determine whether potential matches were definite matches, definite nonmatches or still indeterminate. For example, patients were identified as definite matches if, in addition to the same soundex code and date of birth, their HIV-positive date matched to within the same calendar year, their dates of deaths matched to within 1 month, they were known to have transferred to or from the other centre, the date the patient was last seen at one of the centres was within 1 month of the date that they were first seen at the other centre, or if their country of birth was the same and was not the UK. Indeterminate matches were defined as cases where records came from the same centre, where the individuals were of different sex or had different first initials, or where no other rules were met. Of 1557 records matching on soundex and date of birth, 1150 were determined by the algorithm to be matches. The remaining 407 were deemed indeterminate by the algorithm and were checked manually by two of the investigators – as a result, 155 (38%), 45 (11%) and 207 (51%) were determined to be matches, nonmatches and indeterminate, respectively.

Since soundex codes and dates of birth are prone to errors of transcription and soundexes can be miscoded or derived from false names, the original matching process was widened to include those where date of birth and first-name initial or soundex initial matched, where initials matched but in the reverse order, or where day and month of date of birth matched but the year differed. From this second matching process, a total of 2824 potential matches were identified. Of these, 291 (10%) were determined by the computer algorithm to be definite matches and 2533 (90%) to be indeterminate matches. To confirm the reliability of the algorithm, all of the definite matches and a random sample of 117 of the indeterminate matches were checked manually by two investigators, who then determined whether or not they matched on the basis of all available information. Where the algorithm produced an indeterminate match, the manual checking determined that 80% were nonmatches, 19% were still indeterminate and $<1\%$ matched. Therefore it was felt that in this situation the indeterminate matches produced by the algorithm could be accepted as nonmatches, and no further checking of these would be made. By contrast, where the algorithm produced a positive match, the manual checking determined that only 39% matched, 30% were nonmatches and 31% were still indeterminate. Manual checking was thus thought to be necessary in those cases where soundex and date of birth matched exactly, but where the algorithm produced an indeterminate match, and in those cases where soundex and date of birth did not match exactly but where the algorithm produced a match. When the status of the

indeterminate matches could not be ascertained after manual checking (in most cases insufficient information was available to determine whether the two records were indeed matches) these records were left as distinct individuals in the final dataset.

As a result of these procedures, the cohort contains 1185 merged records (8.6% of patients in the cohort) for individuals considered to have attended two or more centres, often for apparently overlapping periods. Of these 1185 patients, 1078 (91.1%) were known to have attended two of the centres only, 95 (8.0%) had attended three of the centres and 12 (1.0%) had attended four different centres.

When two or more data records from different centres were determined to be related to the same individual, these records were merged. Data merging was accomplished with the development of a computer program that updated any missing information in one record with information from the matched record, and ensured that the merged information reflected the earliest HIV-positive and first-seen dates, and the latest HIV-negative and last-seen dates. Inconsistencies in treatment history merged from different centres were resolved manually by the investigators.

Statistical methods

Date of first entry to the cohort was defined as the date of first attendance at one of the clinics, the date of the patient's 16th birthday or 1 January 1996, whichever occurred latest. Patients were then classified as under follow-up in each year following this date if first-seen and last-seen dates at any of the centres suggested that they were under follow-up at that centre at any time in the year. Thus, if patients transferred to another centre out of the collaboration, they were removed from the risk set from the year following their transfer. If patients subsequently re-attended one of the centres, then they were brought back into the risk set accordingly.

The majority of the data downloads occurred over a 6-month period between October 2001 and March 2002. Whilst data from 2002 are known to be incomplete, data from 2001 are believed to be reasonably complete for most cohorts (although delays in entering new patients on to computerized databases mean that some of those newly seen at a centre in 2001 may not yet be included). Thus, the results reported in this paper relate to the period 1996–2001.

Results

The cohort currently contains information on 13 833 individuals from six of the seven centres (Table 1). As planned at the outset of the study, the seventh centre will submit data at a subsequent data download. Eighty-two per cent of the cohort are male, with an overall median age of 34 years at cohort entry. The predominant risk factor for HIV infection is reported to be homo/bisexual sex (63%) followed by heterosexual sex (24%), with only 4% reporting injection drug use. Fifty-six per cent of the individuals in the cohort are white (ethnicity is unknown for 17%), 26% are known to have developed AIDS, and 8% are known to have died, although the latter may be underestimated as a result of loss to follow-up of some individuals and under-reporting of some AIDS events.

Table 1 Characteristics of patients included in the cohort

Characteristic	*n*	(%)
Total number of patients	13 833	(100.0)
Male	11 364	(82.2)
Female	2469	(17.8)
Ethnicity		
White	7718	(55.8)
Black African	2144	(15.5)
Black unspecified/other	559	(4.0)
Other/mixed	1011	(7.3)
Not known	2401	(17.4)
HIV exposure group		
Homo/bisexual	8682	(62.8)
Injecting drug user	601	(4.3)
Heterosexual	3291	(23.8)
Blood products	103	(0.7)
Other/not known	1156	(8.3)
Country of birth		
UK	5130	(37.1)
Non-UK	3461	(25.0)
Not known	3922	(28.4)
Clinics attended[a]		
Chelsea and Westminster Hospital	5190	(37.5)
Brighton and Sussex University Hospital	1054	(7.6)
King's College Hospital	1516	(11.0)
Mortimer Market	2803	(20.3)
Royal Free Hospital	1991	(14.4)
St. Mary's Hospital	2583	(18.7)
Known to have died	1087	(7.9)
Known to have developed AIDS	3461	(25.0)
Date of first HIV-positive test: median (range)[b]	June 1995	(January 1980–October 2002)
Date of first entry into cohort: median (range)	September 1996	(January 1996–February 2002)
Age (years) at first entry into cohort: median (range)	34	(16–83)

[a]Clinics attended at any time during study period; patients may have attended more than one of the six centres.
[b]First positive HIV test date known for 13 270 patients (95.7%) in the cohort.

Results of data audit

The results of the data audit are summarized in Table 2. In general, the demographic data were considered to be

Table 2 Summary of main results of data audit

	Cohort					
	1	**2**	**3**	**4**	**5**	**6**
Number of records audited	20	27	16	16	26	39
Demographic data						
Date of birth: *n* (%) matching exactly	18 (90)	27 (100)	16 (100)	16 (100)	26 (100)	39 (100)
Soundex: *n* (%) matching exactly	19 (95)	26 (96)	15 (94)	16 (100)	25 (96)	35 (90)
First initial: *n* (%) matching exactly	20 (100)	27 (100)	16 (100)	16 (100)	26 (100)	38 (97)
Sex: *n* (%) matching exactly	20 (100)	27 (100)	16 (100)	16 (100)	26 (100)	39 (100)
Exposure group: *n* (%) matching exactly	19 (95)	26 (96)	12 (75)	15 (94)	26 (100)	37 (95)
Ethnicity: *n* (%) matching exactly	16 (80)	21 (78)	–	9 (56)	25 (96)	39 (100)
Country of origin: *n* (%) matching exactly	–	20 (74)	14 (88)	8 (50)	25 (96)	37 (95)
First-seen date: *n* (%) matching within 1 month	20 (100)	23 (85)	16 (100)	16 (100)	24 (92)	37 (95)
First HIV-positive date: *n* (%) matching within 1 year	19 (95)	26 (96)	16 (100)	16 (100)	26 (100)	35 (90)
Last HIV-negative date: *n* (%) matching within 1 year	20 (100)	26 (96)	15 (94)	13 (81)	22 (85)	33 (85)
AIDS events						
Number of patients with AIDS diagnosis in notes	3	9	4	2	4	14
Number of patients with AIDS diagnosis in database	3	8	3	1	2	11
n (%) in whom AIDS status was correctly identified	20 (100)	22 (81)	15 (94)	15 (94)	24 (92)	36 (92)
Total number of AIDS events in either notes or database	6	12	6	2	6	21
Number of AIDS events in notes but not database	1	3	1	1	3	8
Number of AIDS events in database but not notes *n* (%) of AIDS patients with discrepancy in either date or event:	2	2	0	0	0	0
Of initial AIDS event	1 (5)	3 (14)	1 (7)	1 (7)	2 (8)	3 (8)
Of subsequent AIDS events	2 (10)	2 (9)	0 (0)	0 (0)	1 (4)	5 (14)
Antiretroviral history						
Number of distinct drug 'records' identified from notes or database	60	76	78	25	112	222
n (%) of drug records with discrepancy in date or type of drug	7 (12)	29 (38)	8 (10)	3 (12)	16 (14)	36 (16)

–, data not recorded on clinic database.

reasonably accurate. Whilst data from centres contained inaccuracies in up to 10% of records for a few data items, the other items were accurate. HIV-positive dates were >90% accurate at all centres (note that some centres routinely perform a confirmatory HIV test when a patient first attends, and this date rather than a first-ever positive test date may be the date entered on clinic databases). HIV-negative dates are less commonly recorded, but where available were reasonably accurate. More than 90% of records at centres were accurate for HIV exposure group, and >80% accurate for ethnicity. Where data on exposure group, ethnicity and country of origin were missing from databases, they could often be found in patients' notes. First-seen dates in the database and patients' notes matched to within 1 month for >92% records at five of the six centres (85% at one centre).

One area of potential concern was AIDS events. The AIDS status (whether the patient did or did not have a diagnosis of AIDS) of patients was accurate for 92% or more of patients at five of the six centres, with the database generally missing a couple of patients with an AIDS diagnosis at each centre. Approximately 32% of all AIDS events were found in notes but were not recorded on the database, and 7.5% of AIDS events were recorded on the database but not in the notes. In general, inaccuracies tended to occur in both first and subsequent AIDS-defining events. Where the initial AIDS event did not match, this did not appear to be restricted to any particular type of AIDS event. However, of the eight subsequent AIDS events that did not match, five of these were oesophageal candidiasis. When the records of patients with an AIDS diagnosis were examined in more detail, 5–14% of patients were found to have a record with a mismatch in an initial AIDS event, and 0–14% had a mismatch in subsequent AIDS events.

Of the patient records examined at five of the centres, 84–90% of drug records matched. In the remaining centre, where treatment data recorded on the database were obtained directly from electronic pharmacy prescription records, a much lower percentage of records were judged to be accurate (62%). Subsequently, this centre has arranged to update their clinic database with treatment history taken directly from patients' notes. It should be noted that, in most situations, only one or two drug records in an individual's ART history did not match, rather than the complete history being inaccurate. During the early development of the cohort, the collaborators acknowledged that there could be instances where patients transferred to another centre and their full ART history was not adequately recorded on the new clinic's database. Out of a total of 144 records audited, six patients (4%) at two of the centres were found to have treatment history recorded in the notes but not on the database.

Changing characteristics of patients under follow-up/ newly diagnosed

The number of individuals in the cohort under follow-up has risen each year, from 7294 (53% of cohort) in 1996 to 9375 (67%) in 2000 (Table 3). Between 1996 and 2001, in line with changes in the HIV epidemic in the UK, the percentage of those under follow-up who were male decreased from 87.5 to 82.5%. The proportion of individuals under follow-up with a known homo/bisexual risk for infection decreased by 6%, whilst there was an increase in the proportion of individuals infected through heterosexual sex (8%). The proportion of prevalent cases that were of Black African ethnicity increased from 9% to 17% over the same period. Although loss to follow-up is difficult to define within a clinical setting such as this, between 7.3 and 10.8% of patients in each year did not reappear in the cohort in subsequent years.

Whilst the majority of the cohort were diagnosed as HIV positive prior to 1996, around 1000 new diagnoses have been made in each year since (Table 4). The proportion of newly diagnosed individuals who were female increased from 13% prior to 1996 to 28% in 2001. Over the same

Table 3 Characteristics of cohort under follow-up in each calendar year

			Risk group			Ethnicity		
Year of follow-up	Total	Female *n* (%)	Homo/bisexual *n* (%)	Injecting drug user *n* (%)	Heterosexual *n* (%)	White *n* (%)	Black African *n* (%)	No. (%) lost to follow-up[a]
1996	7294	915 (12.5)	5301 (72.7)	388 (5.3)	1170 (16.0)	4500 (61.7)	653 (9.0)	656 (9.0)
1997	7708	1034 (13.4)	5466 (70.9)	385 (5.0)	1362 (17.7)	4820 (62.5)	799 (10.4)	563 (7.3)
1998	8350	1210 (14.4)	5841 (70.0)	402 (4.8)	1620 (19.4)	5195 (62.2)	995 (11.9)	611 (7.3)
1999	8899	1382 (15.5)	6115 (68.7)	375 (4.2)	1910 (21.5)	5456 (61.3)	1212 (13.6)	711 (8.0)
2000	9375	1566 (16.7)	6294 (67.1)	355 (3.8)	2194 (23.4)	5612 (59.9)	1469 (15.7)	1012 (10.8)
2001	9231	1611 (17.5)	6129 (66.4)	311 (3.4)	2221 (24.1)	5359 (58.1)	1560 (16.9)	–

[a]Defined as any patient under follow-up in current year who did not attend for care at one of the centres in the next or subsequent years.

Table 4 Characteristics of cohort individuals by year of HIV diagnosis

			Risk group			Ethnicity				
Year of follow-up	Total	Female *n* (%)	Homo/ bisexual *n* (%)	Injecting drug user *n* (%)	Heterosexual *n* (%)	White *n* (%)	Black African *n* (%)	Age at diagnosis (years) Median (IQR)	CD4 at diagnosis (cells/μL) Median (IQR)	Viral load at diagnosis (log_{10} HIV-1 RNA copies/mL) Median (IQR)
<1996	7079	910 (12.9)	5112 (72.2)	454 (6.4)	1182 (16.7)	4428 (59.4)	581 (8.2)	29 (25–35)	370 (210–540)	n/a
1996	1122	184 (16.4)	749 (66.8)	33 (2.9)	262 (23.4)	657 (58.6)	165 (14.7)	32 (27–37)	344 (170–530)	4.3 (3.4–4.9)
1997	1075	216 (20.1)	665 (61.9)	31 (2.9)	313 (29.1)	591 (55.0)	199 (18.5)	32 (28–37)	317 (150–513)	4.5 (3.6–5.2)
1998	1069	243 (22.7)	610 (57.1)	30 (2.8)	368 (34.4)	533 (49.9)	251 (23.5)	33 (28–38)	321 (151–487)	4.4 (3.6–5.1)
1999	1030	249 (24.2)	578 (56. 1)	17 (1.7)	382 (37.1)	502 (48.7)	258 (25.0)	34 (29–39)	311 (130–505)	4.5 (3.7–5.2)
2000	1052	283 (26.9)	539 (51.2)	17 (1.6)	420 (39.9)	442 (42.0)	317 (30.1)	33 (28–39)	321 (150–518)	4.6 (3.8–5.2)
2001	786	216 (27.5)	359 (45.7)	11 (1.4)	307 (39.1)	303 (38.5)	242 (30.8)	33 (29–39)	324 (159–521)	4.5 (3.8–5.1)

IQR, interquartile range.

period, the proportion infected via heterosexual sex increased from 17 to 39%, whereas the proportions of new diagnoses that were among homo/bisexual men or injecting drug users decreased accordingly. The proportion of new diagnoses that were in individuals of Black African ethnicity increased from 15 to 31%. Although there were significant differences ($P<0.0001$ in each case) between those newly diagnosed in each year in terms of age at diagnosis, and first CD4 count and viral load after diagnosis, these differences were generally most notable between the groups diagnosed in or before 1996, and those diagnosed subsequently. After excluding those diagnosed in 1996 or earlier, there were no significant differences in either the CD4 count ($P = 0.65$) or viral load ($P = 0.38$) over time, although individuals diagnosed more recently remained slightly older than those diagnosed earlier ($P = 0.007$).

Discussion

The UK CHIC Study is a collaboration of seven of the largest HIV treatment centres in the UK. To date, the cohort contains information on almost 14 000 individuals infected with HIV, including many infected via heterosexual sex, female patients and those from ethnic minority backgrounds. This paper provides information on the development of the cohort and the characteristics of the patients currently enrolled. Future planned analyses of the cohort will describe the uptake of HAART, and will identify factors associated with response to HAART.

When the UK CHIC Study was initiated, a set of data items was agreed upon which included all essential demographics, information on clinical AIDS events, CD4 counts, viral loads, and antiretroviral treatment start and stop dates. It has, however, become apparent that collection of further information (including data on hepatitis B and C status, adherence to antiretroviral therapy, the reasons for stopping each treatment regimen, and prophylaxis for *Pneumocystis carinii* pneumonia) would greatly enhance the value of the study. However, at present there is variation in the way that centres collect certain data items. For example, not all centres currently collect adherence data electronically, and of those that do, the data cover different retrospective time periods and vary in completeness. One advantage of being part of a large multicentre study, such as this, is that centres starting to collect information can benefit from the experiences of those that already record these data on their databases. Centres can thus be encouraged to move towards a more consistent approach towards data collection, not only between UK centres but also with other international HIV datasets.

The quality of the final database is only as good as that of the source databases. To assess the accuracy and completeness of the data at each centre, a sample of 1% of patient records was audited. The approach taken to auditing other HIV cohorts varies greatly, from studies auditing a small sample of all records [14] to others performing complete audits of selected centres within their study [15]. In our case, a small audit sample was felt to be reasonable as many of the cohorts had already published extensively on their own cohorts, and some regularly carry out their own audit procedures. Despite this, the audit did reveal some areas of data collection that could be improved. In particular, AIDS events (particularly subsequent events which are not reported to the HPA-CDSC) and ART data sometimes contained errors. There are a number of possible reasons why the reporting of AIDS events may be poor. In some centres, information collected as part of inpatient visits may be kept separately to that collected at outpatient visits for confidentiality reasons. It is possible therefore that information on AIDS events that occur during an inpatient attendance may not be transferred to the database. Furthermore, AIDS events that occurred before a patient had registered with a clinic may not always

be available (although this information is usually requested by the clinic when a patient first attends). We are already collaborating with the HPA-CDSC to assess the reporting of AIDS events both in this cohort and in the national HIV/AIDS Patient database, and results from this study will be reported separately. The one centre with the highest rate of errors in ART data was the one that had obtained the information solely from electronic pharmacy prescription records; whilst this approach is likely to provide a complete listing of all drugs prescribed to a particular patient *at that centre*, it is likely to miss any treatment history prior to the patient attending the centre (which may be written in notes) and start dates may also be inaccurate if a patient collects a prescription but chooses not to start the drug for a period of time.

Whilst attempts are currently underway to improve the recording of these data (and the data will be re-audited each time a new data download is obtained to assess whether improvements have been made), it was recognized that a certain level of inconsistencies may be unavoidable as a result of differing methods of data collection and entry. For example, demographic data are often collected by clinics on 'first visit' forms that are then used to enter data on to clinic databases. If these forms are not fully completed at the time of the patient's first visit, the data will be missing in the cohort database, even though this information may be subsequently written in patients' notes. Furthermore, any changes made to the patient demographics after the information on the first-visit form has been entered on to the database may not be transferred on to the database. The identification of limitations, such as these, in a source database will enable centres to focus on specific data items and to improve the way in which these are recorded. Any improvement in data accuracy and completeness will be advantageous not only for the larger cohort, but also for centres that use their clinic database as a resource for clinical management or epidemiological research.

As the initial data downloads from centres occurred over a 6-month period, there is variation in the date at which data from the different centres are complete. In particular, the date of last follow-up of patients in each cohort is dictated by the protocols used within the individual centres for updating their own databases. Whilst most cohort data are complete to the end of 2001, we are aware that those centres who submitted data early would naturally have tended to provide data that were less up-to-date than those who submitted data later. Thus, we believe that the number of patients under follow-up in 2001 is an underestimate and that some newly diagnosed patients or recent treatment changes may not have been captured. It is anticipated that subsequent downloads will occur over a much shorter time period, which should reduce this problem.

The large UK CHIC dataset compiled so far is comprised of data for patients who have all attended centres in London and the south-east of England. The data are generally representative of HIV infections in the UK as the demographic data compare well with those of the 2001 Survey of Prevalent Diagnosed HIV Infections (SOPHID) data for the UK [1]. In the UK CHIC Study, there is a slightly higher overall percentage of HIV infections acquired by sex between men compared with SOPHID (63 vs. 54%), which is likely to be a reflection of the patient population in the south-east compared with that of the UK as a whole. The trend in the increasing number of new HIV diagnoses that occur among heterosexuals reported for the UK is also seen in our study, and, of the patients with heterosexually acquired HIV infection under follow-up in 2001, 60% were of Black African ethnicity, compared with 64% reported for the UK overall [1]. We do not believe that any minor differences between the characteristics of the UK CHIC Study and the population of HIV-infected individuals in the UK will be large enough to impact on the results of analyses. However, it is hoped that, in future, the value of this cohort will be increased through expansion to include other UK centres, which would also allow an increased participation in epidemiological research for those centres.

It was considered important to identify and link data for individuals who had attended more than one centre, to ensure that the denominator of patients under follow-up was as accurate as possible. Multiple records relating to one individual could lead to misleading consequences if, for example, treatment data were absent from one record, suggesting that the patient was treatment naïve, but present in another. The fact that data were anonymized meant that identification of duplicates (which was based on soundex code, date of birth and other clinical information only) was often difficult, particularly when data items were missing, when dates were approximated, or when there was very little clinical information available on one or other record. In such situations, where insufficient evidence was available to decide whether or not the two records were indeed for the same individual, the decision was made to leave the records as two separate individuals. Thus, it is possible that, in a small number of cases, records have either been incorrectly linked or been incorrectly left as two separate individuals. It should be reiterated that data were always kept anonymized, and that centres were not informed if their patients may have attended elsewhere for care.

The UK CHIC Study has links to another database currently in development, the National HIV Resistance Database [15]. This database collects results of resistance

tests, including mutations detected in the reverse transcriptase and protease genes and predictions of drug susceptibilities. Many individuals with records in the UK CHIC Study will have had a resistance test during their care, and data from these tests, and from other centres outside the cohort, are being collected. The collection of treatment history together with clinical data is essential for analyses relating to the development of viral resistance. Accordingly, agreement was obtained from the UK CHIC Steering Committee that data will be made available for use with the Resistance database.

As a large multicentre cohort, the UK CHIC Study cohort has similar characteristics to, and is representative of, the HIV-infected population in the UK. The methodology applied in the development of the cohort has made it possible to enhance the accuracy of the data, which have been audited, and to identify and merge data from multiple records relating to the same individuals. Since regular updates of clinical data are planned, the UK CHIC Study will be a useful tool for facilitating research that involves long-term follow-up of groups of patients, or that may lead to the improvement of patient care. The cohort will be of help in determining the feasibility of future studies or clinical trials, as it will be possible to ascertain the size of eligible patient populations. At present, this database is a valuable asset for those investigating the pathogenesis of HIV, opportunistic infections, immunological markers of disease and the impact of antiretroviral treatment in the UK.

Acknowledgements

We would like to thank all the clinicians, data managers and research nurses in participating clinical centres (Appendix 1) who have assisted with the provision of data for this project. This work was funded by the Medical Research Council, UK (Grant G0000199).

References

1 PHLS CDSC, ICH (London), SCIEH. *HIV and AIDS in the United Kingdom 2001. An update November 2002*. London, CDSC, 2002.

2 Layne SP, Marr TG, Colgate SA, Hyman JM, Stanley EA. The need for national HIV databases. *Nature* 1988; **333**: 511–512.

3 Lee JY. Uses of clinical databases. *Am J Med Sci* 1994; **308**: 58–62.

4 Easterbrook PJ. Research potentials and pitfalls in the use of an HIV clinical database: Chelsea and Westminster Hospital. *J Acquir Imm Defic Syndr* 1998; **17** (Suppl. 1): S28–S33.

5 Poulton MB, Sabin CA, Fisher M. Immunological changes during treatment interruptions: risk factors and clinical sequelae. *AIDS* 2003; **17**: 126–128.

6 Mackie N, Sabin CA, Weston R, Weber J. Durability and tolerability of nevirapine (NVP)-containing regimens in a cohort of antiretroviral-naive HIV-positive patients. *HIV Med* 2002; **3**: 172 [Abstract P8].

7 Mocroft A, Sabin CA, Youle L *et al*. Changing treatment patterns among patients with HIV. Royal Free Hospital 1987–97. *HIV Med* 1999; **1**: 32–39.

8 Moyle GJ, Datta D, Mandalia S, Morlese J, Asboe D, Gazzard BG. Hyperlactataemia and lactic acidosis during antiretroviral therapy: relevance, reproducibility and possible risk factors. *AIDS* 2002; **16**: 1341–1349.

9 Mocroft A, Youle M, Morcinek J *et al*. Survival after diagnosis of AIDS: a prospective observational study of 2625 patients. Royal Free/Chelsea and Westminster Hospitals Collaborative Group. *BMJ* 1997; **314**: 409–413.

10 Matthews GV, Sabin CA, Mandalia S *et al*. Virological suppression at 6 months is related to choice of initial regimen in antiretroviral-naive patients: a cohort study. *AIDS* 2002; **16**: 53–61.

11 Easterbrook PJ, Ives N, Waters A *et al*. The natural history and clinical significance of intermittent viraemia in patients with initial viral suppression to <400 copies/ml. *AIDS* 2002; **16**: 1521–1527.

12 Mullen J, Leech S, O'Shea S *et al*. Antiretroviral drug resistance among HIV-1 infected children failing treatment. *J Med Virol* 2002; **68**: 299–304.

13 Mortimer JY, Salithiel JA. 'Soundex' codes of surnames provide confidentiality and accuracy in a national HIV database. *Commun Dis Rep CDR Rev* 1995; **5**: R183–R186.

14 Sudre P, Rickenbach M, Taffe P *et al*. Clinical epidemiology and research on HIV infection in Switzerland: the Swiss HIV Cohort Study 1988–2000. *Schweiz Med Wochenschr* 2000; **130**: 1493–1500.

15 Cohen CJ, Iwane MK, Palensky JB *et al*. A national HIV community cohort: design, baseline, and follow-up of the AmFAR Observational Database. *J Clin Epidemiol* 1998; **51**: 779–793.

16 Dunn D, Matthias R, Hill T, UK Collaborative Group on HIV Drug Resistance. The UK HIV Drug Resistance Database. *HIV Med* 2003; **4**: 209 [Abstract P4].

Appendix 1

The UK CHIC Steering Committee/Writing Committee

Abdel Babiker, David Dunn, Philippa Easterbrook, Martin Fisher, Richard Gilson, Margaret Johnson, Janet Mortimer, Barry Peters, Andrew Phillips, Kholoud Porter, Caroline Sabin, George Scullard, Brian Gazzard (Chair), Ryanne Matthias and Teresa Hill.

Data management group

David Dunn, Philippa Easterbrook, Richard Gilson, Teresa Hill, Ryanne Matthias, Janet Mortimer, Andrew Phillips, Kholoud Porter and Caroline Sabin.

Participating centres

Medical Research Council Clinical Trials Unit (MRC CTU), London (Abdel Babiker, David Dunn, Ryanne Matthias and Kholoud Porter).
King's College Hospital, London (Philippa Easterbrook, Anele Waters, Dorian Crates and Natasha Morgan).
Brighton and Sussex University Hospitals NHS Trust (Martin Fisher, Nicky Perry, Anthony Pullin, Duncan Churchill and Wendy Harris).
Chelsea and Westminster NHS Trust, London (Brian Gazzard, Steve Bulbeck and Sundhiya Mandalia).
Mortimer Market Centre, Camden Primary Care NHS Trust and Royal Free and University College Medical School (RFUCMS) (Richard Gilson, Julie Dodds, Nina Fudge, Andy Rider and Ian Williams).
Health Protection Agency-Communicable Disease Surveillance Centre (HPA-CDSC), London (Janet Mortimer).
St. Thomas' Hospital, London (Barry Peters, Nick Larbalestier and Kimberly Gray).
Royal Free NHS Trust and RFUCMS, London (Teresa Hill, Andrew Phillips, Caroline Sabin, Margaret Johnson, Mike Youle, Fiona Lampe, Colette Smith, Helen Gumley and Clinton Chaloner).
St. Mary's Hospital, London (George Scullard, Jonathan Weber, John Clarke and Christine Owens).