Short communication

HIV cohort collaborations: proposal for harmonization of data exchange

*Jesper Kjær[1] and Bruno Ledergerber[2]**

[1]Copenhagen HIV Programme, Hvidovre University Hospital, Hvidovre, Denmark
[2]Division of Infectious Diseases, University Hospital, Zurich, Switzerland

*Corresponding author: Tel: +41 1 255 3357; Fax: +41 1 255 3291; E-mail: infled@usz.unizh.ch

HIV cohort studies have provided useful information on the natural history of HIV infection and the effects of antiretroviral therapy. It has become increasingly common to combine data from several cohorts into one dataset in order to address certain specific questions with more statistical power than can be achieved with the individual studies. This requires each cohort to map data into a standard format before merging. Until recently, this standard format has differed for each such collaborative analysis. We have therefore developed the HIV Cohort Data Exchange Protocol (HICDEP), which is freely available at http://www.cphiv.dk/HICDEP.pdf. Once individual cohorts have set up a means of transfering data into this format, as and when required, this should greatly facilitate data merging for future joint analyses. The HICDEP incorporates data from HIV drug resistance tests, which have been particularly challenging for cohorts to integrate into databases.

Introduction

Cohort collaborations have been increasingly successful in addressing questions for which individual cohorts do not have sufficient power to provide answers within a reasonable time [1–7]. However, each of these collaborations used proprietary protocols for data exchange requiring substantial data-management efforts, which are potentially error-prone. In addition, cohorts are challenged with the need to exchange information on HIV-1 resistance tests. The aim of this report is to present a recently developed protocol for data exchange between HIV cohorts to facilitate and harmonize future collaborations. This protocol may also provide guidance for the data-management of new cohorts being created.

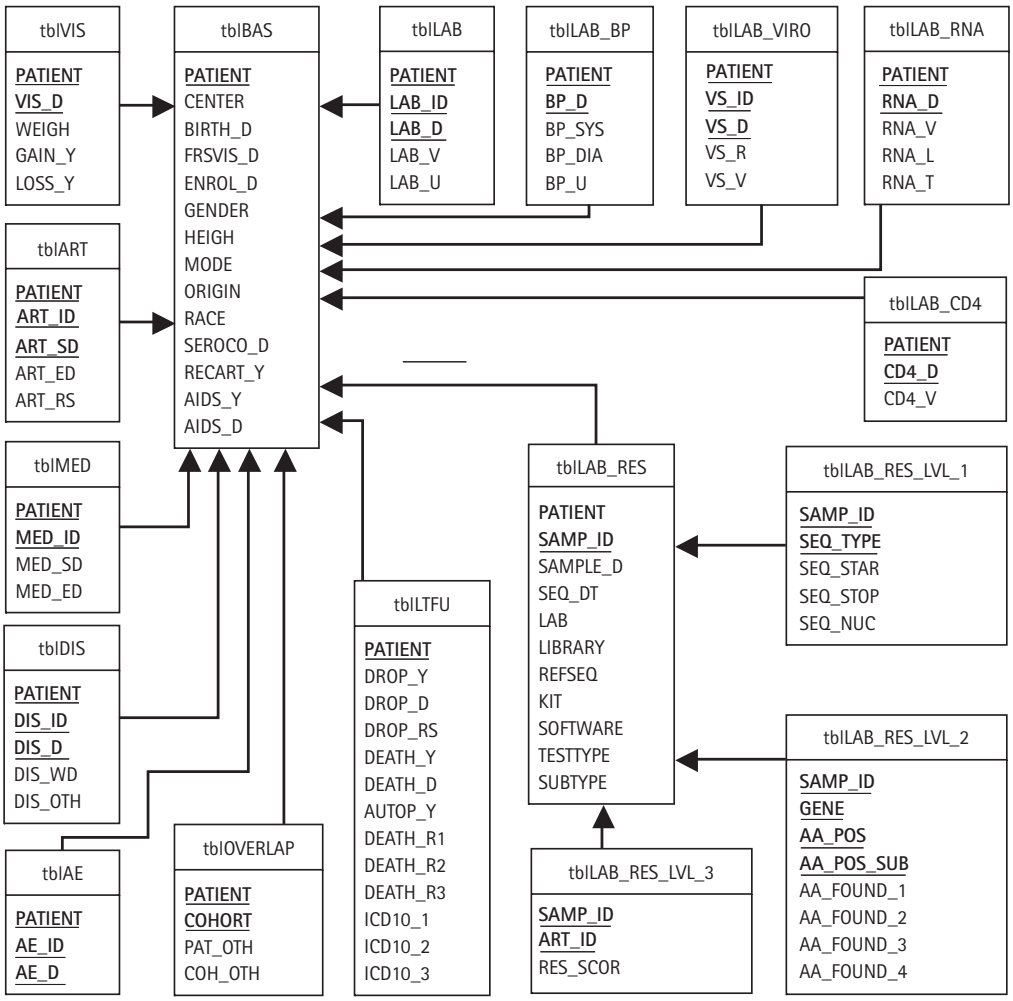
Methods

We mapped the existing data structures of the Data Collection on Adverse Events of Anti-HIV Drugs Study (DAD) [5], EuroSIDA [8] and the Swiss HIV Cohort Study [9] to a relational database. We also evaluated which of the proprietary codes for drugs, diseases and other categorical information, which are individually defined by the various cohorts, could be replaced by common, already published coding systems. Special emphasis was given to the different ways resistance data are collected and reported in research settings and routine clinical practice: for genotypic analyses we considered nucleotide or amino-acid sequences (level 1), amino-acid changes (mutations) from a reference strain (level 2) and drug-specific resistance scores (level 3). Results of phenotypic analyses are also recorded at level 3 together with additional specifications of the methodology used.

Results

The initial versions of the HIV Cohort Data Exchange Protocol (HICDEP) were presented, discussed and amended at the *7th* and *8th International Workshops on HIV Observational Databases*, 28–30 March 2003 in Fiuggi, Italy, and 25–28 March 2004 in Montreux, Switzerland. The main data structure of HICDEP is shown in Figure 1 and currently contains 17 data tables and 18 lookup tables for coding of categorical variables (drugs, reasons for stopping antiretroviral drugs, HIV-related illnesses, ethnicity, transmission category, etc). We applied WHO's Anatomical Therapeutic Chemical (ATC) codes (http://www.whocc.no/atcddd/) for both antiretroviral drugs and other medication and extended it to incorporate investigational drugs. International Classification of Diseases (ICD-10) codes (http://www.who.int/whosis/icd10/) were used for

Figure 1. Main relational data structure of the HIV Cohorts Data Exchange Protocol (HICDEP V1.1)



All tables are grouped around and linked to tblBAS, the table containing the basic patient information. Not shown are lookup tables which contain the coding lists for HIV-related diseases, drugs etc. Variables in bold face represent the unique key of the respective tables and are used for cross-referencing between the various tables.

adverse events and causes of death. CDC stage C [10] diagnoses of severe opportunistic infections or malignancies, however, do not all map onto the ICD-10 structure and thus were coded with short acronyms.

HICDEP is currently being implemented in two collaborations: (i) the DAD study has recently decided to abandon the flat-file format; the data structure of the 5th data-merger in June 2004 for 11 cohorts and approximately 35 000 patients is based on a subset of HICDEP tables and variables, and (ii) some cohorts collaborating in EuroSIDA have started providing follow-up information according to HICDEP.

In addition, EuroSIDA has put considerable effort into the storage of resistance data according to HICDEP. The data have been collected from two sources: (1) sequencing of virus from plasma samples at central laboratories (Badalona, Spain and Buckinghamshire, UK), and (2) clinical virology reports submitted to the central EuroSIDA coordinating office. From source (1), the nucleotide sequences were transmitted as fasta files (http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml) and imported into a level 1 table (sequences) in the database. The translation into a level 2 table (mutations) for statistical analysis has been automated and includes the following steps: (i) alignment, (ii) parsing to identify differences from HXB2-r reference strain, and (iii) translation of codons that differ into amino-acid mutations. Data are being checked for contamination between sequences over time, the presence of frame shift, stop codons and of M, R, W, Y, S, K and B, D, H, N, V nucleotides.

With source (2), copies from clinical virology reports were manually entered (on average 6/h),

followed by 100% cross-checking (on average 15/h). Currently, the level 2 table contains results from 2354 sequences of which 1484 are from the central laboratories and 870 from clinical reports.

A third source of electronic transfer of resistance from the participating sites in EuroSIDA is currently being investigated.

HICDEP is continuously updated and the most recent version is available from http://www.cphiv.dk/hicdep.pdf. Text files with codes used for the lookup tables and a sample database (Microsoft Access®) are also provided and work is in progress to define the format in XML (XSchema). The system developed for handling resistance data in EuroSIDA is planned to be put into the public domain alongside HICDEP as open source. The protocol contains guidance on additional fields that can be added to the set of core fields and concludes with some considerations regarding database management and additional administrative information.

Discussion

With HICDEP we present, according to the best of our knowledge, the first proposal for data exchange between HIV cohorts, which also includes resistance data. Although still in early stages of development, initial applications of HICDEP are promising. Its success in terms of implementations by cohorts and upcoming collaborations will depend on the timeliness with which new developments in the field are incorporated. One of the next updates will include extended information on the coding of causes of death for which an international harmonization effort is currently under way. We also invite researchers to contact us (jkj@cphiv.dk) regarding improvements or suggestions for extensions (for example, information on adherence or quality of life) they may have experience in.

In conclusion, we are convinced that large-scale cohort collaborations (such as proposed by Brun-Vézinet *et al*. in this issue) will continue to provide crucial answers and harmonization efforts, and, in the long run, will pay off. The HICDEP protocol and complementary files may be especially useful for new cohorts, namely in developing countries, or for those cohorts looking to update their structure.

Acknowledgements

We are grateful to all the colleagues who gave valuable input to the initial versions of the HICDEP protocol, namely Ronan Boulme, Zoé Fox, Nina Friis-Møller, Stephen Hart, Olivia Keiser, Martin Rickenbach, Caroline Sabin, Allen Sawitz, Morten Svenson and Sarah Walker. Special thanks to Andrew Phillips for his helpful comments on this article.

References

1. CASCADE Collaboration. Changes in the uptake of antiretroviral therapy and survival in people with known duration of HIV infection in Europe: results from CASCADE. *HIV Medicine* 2000; **1:**224–231.
2. Phillips AN, Staszewski S, Weber R, Kirk O, Francioli P, Miller V, Vernazza P, Lundgren JD & Ledergerber B. HIV viral load response to antiretroviral therapy according to the baseline CD4 cell count and viral load. *Journal of the American Medical Association* 2001; **286:**2560–2567.
3. Ledergerber B, Mocroft A, Reiss P, Furrer H, Kirk O, Bickel M, Uberti-Foppa C, Pradier C, d'Arminio MA, Schneider MM & Lundgren JD. Discontinuation of secondary prophylaxis against *Pneumocystis carinii* pneumonia in patients with HIV infection who have a response to antiretroviral therapy. Eight European Study Groups. *New England Journal of Medicine* 2001; **344:**168–174.
4. Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, d'Arminio MA, de Wolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B & Sterne JA. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *Lancet* 2002; **360:**119–129.
5. Friis-Moller N, Sabin CA, Weber R, d'Arminio MA, El Sadr WM, Reiss P, Thiebaut R, Morfeldt L, De Wit S, Pradier C, Calvo G, Law MG, Kirk O, Phillips AN & Lundgren JD. Combination antiretroviral therapy and the risk of myocardial infarction. *New England Journal of Medicine* 2003; **349:**1993–2003.
6. Phillips AN, Lundgren JD, Hogg RS, d'Arminio Monforte A, Castelli F, Walker AS, Staszewski S, Dabis F, Gazzard BG, Saag MS, Petoumenos K, Johnson M, Scullard G, Gill MJ, James I, Fisher M, Mussini C, Klein M & Costagliola D. Pre- and within-HAART nucleoside analogue use and viral load rebound on HAART. *Journal of Infectious Diseases* 2004; in press.
7. Ledergerber B, Lundgren JD, Walker AS, Sabin C, Justice A, Reiss P, Mussini C, Wit F, d'Arminio MA, Weber R, Fusco G, Staszewski S, Law M, Hogg R, Lampe F, Gill MJ, Castelli F & Phillips AN. Predictors of trend in CD4-positive T-cell count and mortality among HIV-1-infected individuals with virological failure to all three antiretroviral-drug classes. *Lancet* 2004; **364:**51–62.
8. Kirk O, Mocroft A, Katzenstein TL, Lazzarin A, Antunes F, Francioli P, Brettle RP, Parkin JM, Gonzales-Lahoz J & Lundgren JD. Changes in use of antiretroviral therapy in regions of Europe over time. EuroSIDA Study Group. *AIDS* 1998; **12:**2031–2039.
9. Sudre P, Rickenbach M, Taffe P, Janin P, Volkart AC & Francioli P. Clinical epidemiology and research on HIV infection in Switzerland: the Swiss HIV Cohort Study 1988–2000. *Schweizerische Medizinische Wochenschrift* 2000; **130:**1493–1500.
10. Anon. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Recommendations & Reports: Morbidity & Mortality Weekly Report* 1992; **41:**1–19.

Received 19 May 2004, accepted 21 July 2004