Integrating Phenotypic Data For Depression

Amy W. Butler, Sarah Cohen-Woods, Anne Farmer, Peter McGuffin, Cathryn M. Lewis

Institute of Psychiatry, King's College London, Denmark Hill, London SE5 8AF, U.K.

Abstract

The golden era of molecular genetic research brings about an explosion of phenotypic, genotypic and sequencing data. Building on the common aims to exploit understanding of human diseases, it also opens up an opportunity for scientific communities to share and combine research data. Genome-wide association studies (GWAS) have been widely used to locate genetic variants, which are susceptible for common diseases. In the field of medical genetics, many international collaborative consortiums have been established to conduct meta-analyses of GWAS results and to combine large genotypic data sets to perform mega genetic analyses. Having an integrated phenotype database is significant for exploiting the full potential of extensive genotypic data. In this paper, we aim to share our experience gained from integrating four heterogeneous sets of major depression phenotypic data onto the MySQL platform. These data sets constitute clinical data which had been gathered for various genetic studies for the past decade. We also highlight in this report some generic data handling techniques, the costs and benefits regarding the use of integrated phenotype database within our own institution and under the consortium framework.

1 SUMMARY

As with other human diseases, phenotypic data for psychiatric disorders has been collected for a wide range of clinical and genetic studies across different research centres. Demographic and clinical data sets, either unstructured, semi-structured or structured, are often scattered and held on different systems. Our Unipolar Depression GWAS project genotyped ~3100 cases and ~1700 controls from the samples of three previously conducted studies of recurrent depression and a pharmacogenetic study of antidepressant response. Phenotypic data were used in concert with genotypic data for the core association analyses and as covariates for secondary analyses to identify genetic variants that confer susceptibility to depression. Harmonised genetic and integrative phenotypic data are important to ensure powerful and reliable statistical analyses of local studies, and for participating in collaborative meta-analyses of major depressive disorder (MDD) through the Psychiatric GWAS Consortium (PGC). Building an integrated database for depression phenotype data was an integral part of our MDD GWAS project. Not only allows efficient retrieval of demographic and clinical data, it also provided us with an opportunity to reconcile scattered heterogeneous data sets from different restospective studies for depression. Taking into consideration of data sharing and strategic bioinformatic analytics, the integrated MDD phenotype database was designed with flexibility for easy incorporation of new phenotypic data sets and fitting in with other knowledge base system.

We completed the building of the MDD database on MySQL at the time of this report. It has been fully operational and used as an invaluable resource for both our MDD GWAS and many other local psychiatric research projects (e.g. anxiety, suicidal behaviour, depression linkage follow-up study with more stringent affection status, identifying novel subtypes of depression, comorbidity with BMI and T2D metabolic syndrome study, etc.). Without the

integrated MDD phenotype database, investigators have to mess around with a bunch of Excel, Stata and SPSS files to pull the data together each time they conduct a new study on these data sets.

One of the aims of this report is to give an account of how we structured and integrated the four highly heterogeneous phenotypic data sets, the issues arising from the process of integration from Excel, SPSS and STATA onto MySQL. A discussion on how to meet key requirements of data such as de-identification of subjects and data harmonisation is also included. There are many ways to implement a database project, the purpose of this report is mainly to share our experience with those bioinformaticians who are about to embark on similar data integration projects.

2 INTRODUCTION

The golden era of genetic research results in a big bang of biological and genotypic data. Although it might not have been obvious from the outset, managing vast amounts of biological, genotypic and phenotypic data has become an integral part of most modern day genetic research projects. Genotypic data is comparatively standard in format and structure. Phenotypic data on the other hand may vary greatly in content, structure and degree of data quality. With regard to clinical studies, many psychiatric research projects started many years ago. Longitudinal clinical data collections have been ongoing. Phenotypic data were commonly held in flat files, semi-structured SPSS, STATA or Excel spreadsheets, or simply put together in some Microsoft© Access tables. Proactive data quality measures were often neglected in these systems.

Large sample size provides greater statistical power for GWAS. In order to secure a large number of subjects for the study, investigators often need to combine samples from previous clinical research studies. Stepping up the scale, collaborative efforts aiming to investigate the genetic basis of different human diseases are gathered through establishment of international consortiums for GWAS. Genetic projects under these settings expect shared data, both genotypic and phenotypic, to be harmonised, held and readily accessible in a controlled manner.

At the Institute of Psychiatry, the unipolar depression GWAS (Lewis CM et al. 2010) was designed to draw case and control data from three retrospective studies of recurrent depression and a pharmacogenetic study of antidepressant response. The three depression studies varied considerably in study design, a more detailed description is given in the METHOD section. Our unipolar depression GWAS is one of PGC's contributing studies for MDD. The PGC (https://pgc.unc.edu/) was set up to conduct meta-analyses of GWAS data on five major psychiatric disorders: autism, attention-deficit hyperactivity disorder, bipolar disorder, MDD and schizophrenia. With the involvement of over 80,000 subjects (each has over 500,000 genotyped SNPs) and 40 billion total genotypes, these GWASs constitute the largest genetic research ever conducted in psychiatry. By mapping the PGC's global aims to our local phenotypic data requirements, it implies that: (i) we have to assure that our studies use comparable diagnostic constructs, (ii) our integrated database is able to provide phenotypic data recorded on each subject for those data items that are required for the PGC GWAS analyses, and (iii) all phenotypic data have to be de-identified to ensure source transparency.

3 METHODS / APPROACH

3.1 Data from four depression studies

The first depression phenotypic data set (S1:DeCC/BaCC) was sourced from the Depression Case and Control study. Its data collected from three UK centres: London, Cardiff and Birmingham were held in one flat SPSS file. The second depression data set (S2:DeNt) was originated from the Depression Network study which was based on a family linkage study design (Farmer A et al. 2004). In S2, data on both probands and affected sibling pairs were collected by seven European sites: Aarhus in Denmark, Bonn in Germany, Dublin in Ireland, Lausanne in Switzerland, London; Cardiff and Birmingham in the United Kingdom, and St. Louis in the United States. The source of the third depression data set (S3:GCC) was a depression case and control study with its data collected from two European sites: Bonn in Germany and Lausanne in Switzerland. The phenotypic data of S2 and S3 were semi-structured in Excel files. The fourth data set (S4:GENDEP) belongs to the the Genome-based Therapeutic Drugs for Depression research project (Aitchison KJ et al. 2005). S4 included case subjects who have been investigated for response to treatment with two particular antidepressants. Its data were collected from London and eight other European sites: Aarhus in Denmark, Brescia in Italy, Brussels in Belgium, Ljubjana in Slovenia, Mannheim and Bonn in Germany, Poznan in Poland and Zagreb in Croatia. Unlike the other three studies, the S4 data were overly normalised and held as a relational database in MySQL. However, we found it difficult to dissect the entities with the lack of documentation on its original schema design. Given as an alternative, we were offered a subset of this data in STATA files which provided a clearer picture on the major entities and attributes.

3.2 Road map for the design phase

A road-map of our design and implementation approach is given in Figure 1.

3.2.1 Conceptual modelling phase

We began the MDD phenotype database project with a conceptual analysis. This was accomplished with the expertise knowledge input from our consultant psychiatrists and researchers in the Institute of Psychiatry. A high level conceptual model was drawn to reflect the concept of the subject area and to identify the major entities. We used the conceptual model to define the scope of the integrated MDD phenotype database.

Common across all source studies that provided data for the database, personal demographic data were collected for both case and control individuals. All depression cases had been interviewed using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing JK et al. 1998) focusing on their worst and second worst episodes of depression in the recurrent depression studies, except for S4 which included the current episode of depression only. A variety of other phenotypic measures relating to personality and psycho-social events, together with family medical history were collected under interview or self-report setting.

3.2.2 Logical modelling phase

Based on the conceptual model (Figure 2), we then examined each source data set in turn. We assessed study specific information on data items (e.g. format, domain), local entity keys, data labels, coding schemes for data values; how missing data and unknowns were handled; and how close each data set overlapped with the defined scope of the integrated MDD phenotype database. We conducted a data item mapping exercise to establish what attributes

were available at source for each major entity in the conceptual model. The data mapping process also generated an account of various identifiers used in the source data sets, and a range of coding schemes that we should be aware of when we got to the stage of data harmonisation. The deliverable of this stage was a logical database schema in the form of entity-relationship diagram (ERD). The ERD defined the key and non-key attributes for each identified entity and how the entities were connected to each other via defined relationships. We applied a directory framework to de-identify subject details. This approach requires a new entity to hold all local identifiers and to assign a system generated surrogate key to uniquely identify each subject within the integrated MDD phenotype database. Within the integrated database, phenotypic data on each subject are only linked by the surrogate key (subject_gid). External projects that require de-identified information will only be given the subject_gid to identify individual subject. The local identifiers are reserved specifically for facilitating data population and data change management.

3.2.3 Physical modelling phase

We chose MySQL to be our database management system. MySQL is commonly used for many well known bioinformatic databases, for instance, Ensembl as a joint project of the European Bioinformatics Institute and the Sanger Centre developed their public database servers using MySQL (Birney E et al. 2004), the UCSC Genome Browser set up a MySQL database for public access at "genome-mysql.cse.ucsc.edu" as an alternative to using the table browser interface (Hinrichs AS et al. 2006). Many other good reasons for the popularity of MySQL are listed at the MySQL website (http://www.mysql.com/why-mysql). After the logical database design, we proceeded to produce the physical database schema which was specific for the MySQL platform. We mapped entities to database tables and attributes to table columns. All database objects were defined using data definition SQL scripts. Key attributes were translated into primary keys. We modelled the relationships using foreign key constraints for referential integrity, and created indexes for efficient cross table joins and fast data retrieval.

3.3 Physical implementation

3.3.1 Data migration

Due to the heterogeneity of our source data sets, we used a staging databasing strategy for data migration. Each source dataset was migrated from either SPSS or STATA or Excel into individual MySQL staging database. The SPSS file sourced from the S1 study holds a data set with a dimension of 10,809 variables x 5,587 instances. We had to partition the variables and exported them into manageable chunks. For the STATA files from the S4 study, complication arised from fields of date type for which we had to reformat before the data migration. Using the staging databases as intermediate data store made it easier for data mapping, data transformation and data cleansing because both the staging databases and the integrated phenotype database were held on the same development platform. Elementary data cleansing, mostly detecting and remediating syntactical errors (e.g. typographical errors, invalid dates, misplaced columns, etc.), was carried out within each staging database to ensure data quality was acceptable before the data were integrated.

3.3.2 Data mapping and data population

Before loading the phenotype data into the integrated database, we had to set up the subject directory. This was to ensure each subject from the source data sets would have a unique

surrogate identifier. This was accomplished by defining a global subject identifier in the subject_directory table as "INT UNSIGNED AUTO_INCREMENT PRIMARY KEY". The subject_directory table was designed to hold all local identifiers used at source. Upon the INSERT of a subject instance into the subject_directory table, MySQL would automatically generate a unique identification number (subject_gid) to the new row, and the local identifier(s) that belonged to the subject would also be stored simultaneously. Within the integrated MDD phenotype database, all phenotypic information related to the new subject would only be linked using the subject_gid.

A second wave of data cleansing and missing data recovery were undertaken after the data integration. This time the data cleansing was focused on: (1) consolidating the clinical diagnostic codes and making sure that they conform to the standard format as stated in the ICD-10-DCR (International Classification of Diseases 10th edition, Diagnostic Criteria for Research) and DSM-IV (Diagnostic and Statistical Manual 4th edition operational criteria) manuals. And (2) remediating semantic data errors which could be validated against elementary data item(s) (e.g. age at interview was checked against date of birth and interview date). During the data integration process, we discovered some case subjects had no SCAN data. After we contacted the source data owner, these missing SCAN data were later recovered from a set of Access databases.

3.3.3 Data harmonisation

Harmonised data on attribute level facilitate data management and knowledge mining. Taking the data item of affection status as an instance, it was inconsistently coded across studies (e.g. S1 used '0' for unaffected, '1' for affected and blank for unknown status. S2 used '0' for unknown status, '1' for unaffected and '2' for 'affected'. S3 used 'CASE' for affected status whilst all subjects in S4 were set to default as 'case'). We found several other important attributes (e.g. gender, diagnosis codes) that lacked a harmonious uniformity. To resolve inconsistent data coding schemes, we applied MySQL user defined data type (i.e. ENUM) for data item definitions wherever necessary and appropriate. Basically, an "ENUM" type in MySQL is a string object with a value chosen from a list of allowed values defined by the user. These allowed values are enumerated explicitly in the column specification at table creation time. For example, we defined 'affection_status' as ENUM('affected', 'unaffected', 'unknown') in the subject demographic details table. During data loading from the staging database to the integrated database, we manipulated the mapping and transformed the source data using the "CASE WHEN" clause in the INSERT/SELECT statement, for example:

```
INSERT INTO integratedDB.subj_demographic_det
(subject_gid, affection_status, ...)
SELECT dir.subject_gid,
CASE s2.aff_status WHEN 0 THEN 'unknown'
WHEN 1 THEN 'unaffected'
WHEN 2 THEN 'affected'
END AS 'affection_status',
...
FROM stagingDBs2.demographic_tbl s2, integratedDB.subject_directory dir
```

WHERE s2.patient_id = dir.patient_id;

3.3.4 Meta data

We used meta-data to provide additional information about data labels. In the integrated MDD phenotype database, all SCAN data items and phenotypic measures obtained from various interviewing instruments are referred to by column names (e.g. s01_002 for questionnaire item #002 in SCAN section 1). Although we could use "COMMENT" in the table definitions, the details documented therein would not be accessible to the data users. The meta-data tables were designed to provide a catalogue of descriptions for all interview and self-reported questionnaire items. Clinical diagnoses for case subjects are held as codes compliant to the valid lists published in the ICD-10-DCR and DSM-IV manuals. Similarly, we set up a meta-data table to improve information content so that the users can look up what each abbreviated diagnostic code stands for. The meta-data have added value to the integrated MDD phenotype database as none existed before in any of our source data sets.

3.3.5 Data administration

For housekeeping purpose, each record in the integrated MDD phenotype database has a date-stamp to keep track of last updated date. For development control and change management purposes, we also keep a mirror copy of the operational database as a test bed for testing changes propagated from the source studies. Regular backup of the operational database is carried out by the I.T. support in the institute.

4 RESULTS

We have designed and developed an operational database for the integrated phenotypic data for depression. Not only have we integrated the scattered heterogenous source data sets, we also improved the data quality and added value to the depression phenotypic data. The integrated MDD phenotype database contains 20 tables and holds phenotypic data for over 8,500 subjects. The data have been harmonised to support the GWAS project for unipolar depression. The integrated MDD phenotype database also meets the phenotypic data requirements set out by the PGC for collaborative GWAS meta-analyses. Before we integrated the data, the source data sets were scattered at different physical locations. Unknown to most investigators were the data content and data coverage in each data set, and most significantly the hidden parameter of data quality. Furthermore, most of the depression data were not readily accessible as most scientists are not equipped with I.T. skills to access and merge those disparate data sets from SPSS, Excel, Stata, Access and MySQL. The integrated MDD phenotype database has resolved most of these significant concerns. Its clean data have provided greater power to statistical analyses and increased reliability to scientific hypotheses being tested.

5 DISCUSSION

This database build project has highlighted the need and benefits for integrating scattered phenotypic data sets. Many factors, such as uncertainty regarding data contents and data quality, arising from the heterogeneous source data sets have made it difficult to consider using any of the existing data migration tools or advanced data integration techniques. Handling a flat SPSS file with over ten thousands variables was a great challenge. The data integration was further complicated by differential data handling practices at source. For

instance, some semantically consistent entity properties were represented differently by various number of elementary data items. Using multiple subject identifiers within study was common in most source data sets. Nonetheless, the directory approach for de-identifying subjects has proved to be a workable solution. The integration process enabled us to harmonise a large amount of phenotypic data, this certainly influences the ease of data mining. Having combined the phenotypic data into one integrated database on MySQL, we are able to gather and supply consistent harmonised data to support various research projects for depression or on related endo-phenotypes. This was difficult before the integrated MDD phenotype database was built when researchers had to delve into individual data sets and not knowing what data was available and how to put the data together from heterogeneous systems. The endophenotypes concept has been commonly used to dissect genetics of neuropsychiatric disorders (Gottesman I et al. 2003, Braff DL et al. 2007). Endophenotypes and co-morbidity studies enable classifications by decomposing complex psychiatric disorder diagnoses. There were successful stories about genetic analyses finding endophenotypes meeting association with a candidate gene or genetic loci rather than with the disease syndrom itself. So far, our integrated MDD phenotype database has supported several projects that go down this route.

In the psychiatric research community, we have the growing need to cross search and derive new reliable scientific conclusions from exploring a spectrum of exisitng information resources (e.g. phenotypic and genotypic data, all the ~omics data etc.). It is desirable to have a unified data management infrastructure which is capable to automate linking and aid analysing data hosted by independent sources. If building the integrated MDD phenotype database was part of a substantial bioinformatics project, we would definitely consider more advanced data integration techniques such as intelligent data fusion techniques using resouce description framework, or semantic data interlinking using ontology-based matching and merging (Lopez X et al. 2007). However, this is far beyond the scope and limited by the resource constraint of the current MDD GWAS master project.

We chose not to over model the schema because that would incur excess cost for denormalisation. More importantly our model was aimed for scalability. Using the current integrated schema as the reference model, any new depression phenotype data set can easily be mapped and merged, or interlinked as an extension to the core database. Our semi-normalised schema design also benefits for developing biomart. BioMart is an open source data management system (www.biomart.org) which has evolved to be a generic data integration solution and comes with a range of query interfaces (Smedley D et al. 2009). Institutions can develop and implement their own BioMart system for domain specific querying their own data sources, and at the same time accessing a range of external biomarts using the BioMart central portal architecture (Haider S et al. 2009). The design of our current integrated MDD phenotype database model can easily be modified and made compliant with the reversed star model as specified by BioMart. Adoption of Biomart seems to be the next appropriate step if we want to build a one-stop-shop for integrated resources, which would aid knowledge mining and help further complex analyses in psychiatric disorder research.

Acknowledgements

This study was funded by joint grant from the Medical Research Council, UK and GlaxoSmithKline (G0701420).

References

[1] Lewis CM, Ng MY, Butler AW, Cohen-Woods S, Uher R, Pirlo K, Weale ME, Schosser A, Paredes U, PhD, Rivera-Sanchez M, Craddock N, Owen MJ, Jones L, Jones I, Korszun A, Aitchison KJ, Shi J, Quinn J, MacKenzie A, Vollenweider P, Waeber G, Heath S, Lathrop M, Muglia P, Barnes MR, Whittaker JC, Tozzi F, Holsboer F, Preisig M, Farmer AE, Breen G, Craig IW, and McGuffin P. Genome-wide association study of major recurrent depression in the UK population. AJP-09-09-1380, paper accepted for publication in the Am. J. Of Psychiatry on 28th Jan 2010.

[2] Farmer A, Breen G, Brewster S, Craddock N, Gill M, Korszun A, Maier W, Middleton L, Mors O, Owen M, Perry J, Preisig M, Rietschel M, Reich T, Jones L, Jones I, and McGuffin P. The Depression Network (DeNT) Study: methodology and sociodemographic characteristics of the first 470 affected sibling pairs from a large multi-site linkage genetic study. BMC Psychiatry, 2004, 4:42.

[3] Aitchison KJ, Basu A, McGuffin P and Craig I. Psychiatry and the 'new genetics': hunting for genes for behaviour and drug response. British Journal of Psychiatry 2005, 186: 91-92

[4] Wing JK, Sartorius N and Ustin TB. Diagnosis and clinical measurement in psychiatry. A reference manual for SCAN. World Health Organization, 1998.

[5] Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H-R, Iyer V, Jekosch K, Kahari A, Kasprzyk A, DKeefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, and Clamp M. An Overview of Ensembl. Genome Res. 2004 14: 925-928.

[6] Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, and Kent WJ. The UCSC Genome Browser Database: update 2006. Nucleic Acids Research, 2006, Vol. 34, Database issue D590-D598.

[7] Gottesman II and Gould TD. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. Am J Psychiatry 160:636-645, April 2003.

[8] Braff DL, Freedman R, Schork NJ, and Gottesman II. Deconstructing Schizophrenia: An Overview of the Use of Endophenotypes in Order to Understand a Complex Disorder. Schizophrenia Bulletin 2007 33(1):21-32.

[9] Lopez X and Souripriya D. Semantic Data Integration For the Enterprise. An Oracle White Paper, June 2007.

[10] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, and Kasprzyk A. BioMart – biological queries made easy. BMC Genomics 2009, 10:22.

[11] Syed Haider, Benoit Ballester, Damian Smedley, Junjun Zhang, Peter Rice, and Arek Kasprzyk. BioMart Central Portal—unified access to biological data. Nucleic Acids Research, 2009, Vol. 37, Web Server issue W23–W27.

URLs

GENDEP Official Site: http://gendep.iop.kcl.ac.uk

MySQL home page: http://www.mysql.com

Ensembl public MySQL Servers webpage: http://www.ensembl.org/info/data/mysql.html

BioMart Project home page: http://www.biomart.org

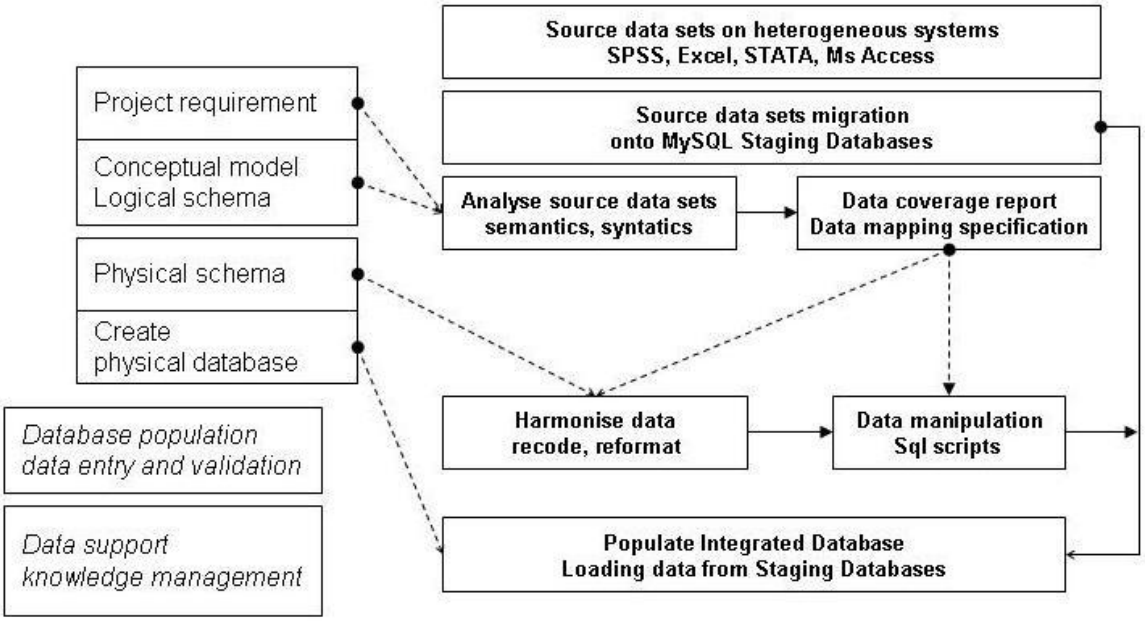

Figure 1: The left-hand side shows the process for top-down design path. The right-hand side shows our current approach, a reverse engineering path as we were committed to use existing data sets rather than populating the database using a pre-designed user interface with data validation procedures incorporated.

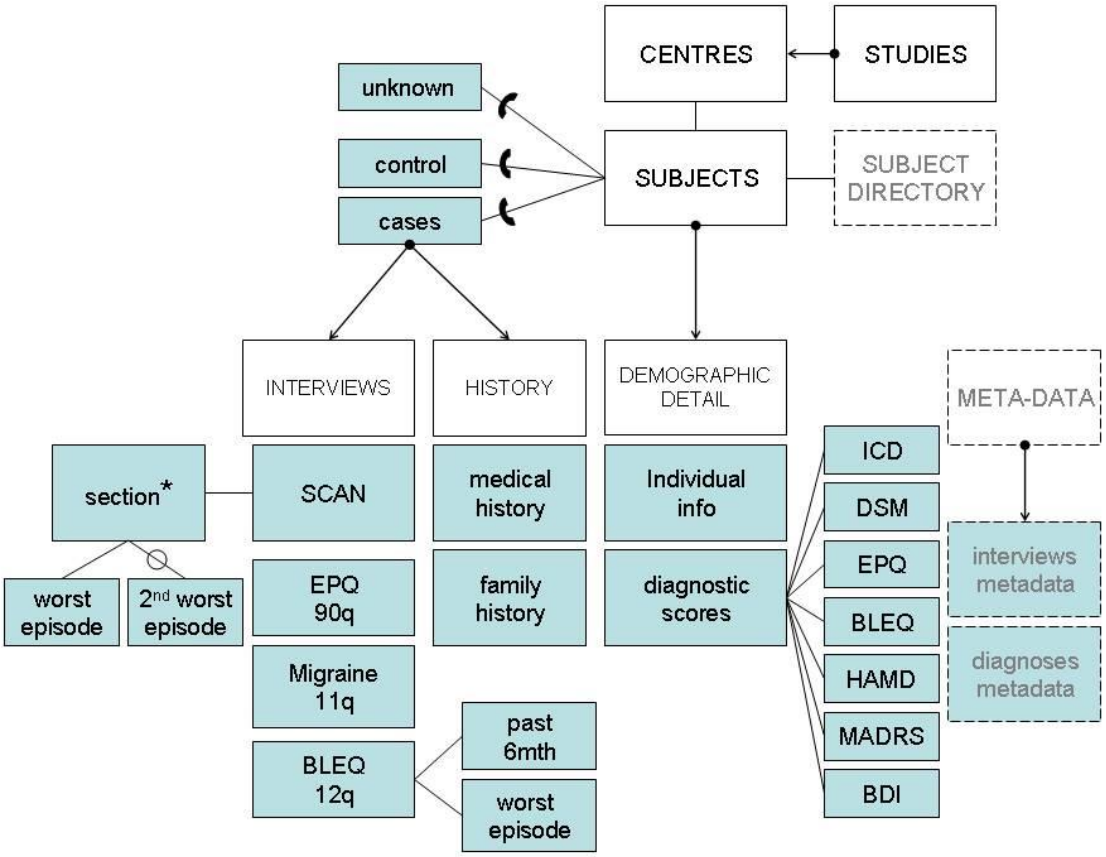

Figure 2: Boxes with solid border represent major entities or key attributes. Lines represent links between entities: optional link with a circle on the line, line with an arc means exclusive-or. Boxes with dash-line border are added concepts for the integrative phenotype database.