IN PRESS

Molecular Psychiatry (2012), 1–6
© 2012 Macmillan Publishers Limited All rights reserved 1359-4184/12
www.nature.com/mp

PERSPECTIVE

Managing sensitive phenotypic data and biomaterial in large-scale collaborative psychiatric genetic research projects: practical considerations

SY Demiroglu[1], D Skrowny[1], M Quade[1], J Schwanke[1], M Budde[2], V Gullatz[2], D Reich-Erkelenz[2], JJ Jakob[1], P Falkai[3], O Rienhoff[1], K Helbing[1], U Heilbronner[2] and TG Schulze[2]

[1]*Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany,* [2]*Section on Psychiatric Genetics, Department of Psychiatry and Psychotherapy, University Medical Center Göttingen, Göttingen, Germany and* [3]*Department of Psychiatry and Psychotherapy, Ludwigs-Maximilians-University, München, Germany*

Large-scale collaborative research will be a hallmark of future psychiatric genetic research. Ideally, both academic and non-academic institutions should be able to participate in such collaborations to allow for the establishment of very large samples in a straightforward manner. Any such endeavor requires an easy-to-implement information technology (IT) framework. Here we present the requirements for a centralized framework and describe how they can be met through a modular IT toolbox.

Molecular Psychiatry (2012) **0,** 000–000. doi:10.1038/mp.2012.11

Keywords: biobanking; data protection; GWAS; IT infrastructure; pseudonyms; sequencing

Psychiatric genetic research has reached a crucial point in time. There is a broad consensus that a variety of methods and approaches will be used jointly. Although genome-wide association studies will not constitute the methodological mainstay anymore, they will continue to be one pillar of gene discovery. They will be complemented by studies of rare variants and copy-number variations. Moreover, whole genome or whole exome sequencing are gaining popularity, with both case-control and family-based designs being applied.[1] A hallmark of current research activities is the emphasis on very large samples, totaling several tens of thousands of individuals, that can only be achieved through multisite collaborative efforts.[2] At present, such multi-site collaborative efforts are very often pursued by joining samples initially collected for stand-alone projects. This not only increases the level of sample heterogeneity but also introduces massive challenges to information technology (IT). Scientists in large multisite projects are increasingly taking advantage of centrally located internet-based administration tools[3] and large repositories for biomaterial, genotype, and phenotype data such as the Rutgers University Cell and DNA Repository (http://www.rucdr.org[4]) or the database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/gap[5]). Particularly in the United States, these repositories and databases are widely used and provide platforms for biomaterial and data sharing. For publicly funded projects, depositing of biomaterial and data may even be mandated by law. Nonetheless, there is a need for highly flexible IT frameworks that allow for ever-growing networks of research groups that differ in size, expertise, research foci, or funding but are joined by the wish to collaborate in the establishment of large samples. The IT challenges that such frameworks have to face are the massive data amount, different data quality, heterogeneous IT sources and, most importantly, nonstandardized metadata. Therefore, multicentric studies in psychiatric genetics should ideally be conceptualized as multicentric from the start. To achieve the needed sample sizes, the research community may in the future need to consider tapping into clinical resources that are not necessarily part of large research institutions. Large samples or even longitudinal cohorts may be drawn from both academic and non-academic centers. Such a framework requires the establishment of flexible IT tools for the deposition of phenotypic data and biomaterial. Given that the technological developments will enable a more and more detailed look on a person's genome, this flexibility will have to be accompanied by highest standards for the protection of sensitive data.[6,7] Although this is true for genetic research on any complex phenotype, psychiatric genetics is typically subject to deeper scrutiny by institutional review

Correspondence: Dr SY Demiroglu, Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Strasse 40, 37075 Göttingen, Germany or Professor TG Schulze, Section on Psychiatric Genetics, Department of Psychiatry and Psychotherapy, University Medical Center Göttingen, Von-Siebold-Strasse 5, 37075 Göttingen, Germany.
E-mail: sara.demiroglu@med.uni-goettingen.de or thomas.schulze@med.uni-goettingen.de

Received 31 October 2011; revised 24 January 2012; accepted 25 January 2012

boards or ethics committees than genetic studies on somatic disorders. Thus, any IT framework for large-scale psychiatric genetic collaborative research has to meet several requirements and challenges:

(a) It should enable easy participation by new sites at any point in time, regardless of their geographical location and sophistication of available IT.
(b) It should allow individual participating sites to control their level of participation in the overall project. This includes the flexible handling of various levels of data and biomaterial data sharing agreements across sites.
(c) It should allow for longitudinal assessment and biobanking, a strategy that has been widely ignored so far, although there is ample evidence that longitudinal studies will broaden our understanding of psychiatric phenotypes and gene-environment interactions.[8–10]
(d) It should adhere to the highest standards of data protection. With the increasing amount of phenomic and genomic data generated from study participants, their risk of being (re-)identified rises.
(e) It should allow for the assessment of a variety of phenotypes (psychological tests, disease history, and so on) and any biomaterial (whole blood, DNA, RNA, proteins, and so on).
(f) It should ensure easy identification and retrieval of samples, and avoid mix-up of samples and lost identity.
(g) It should allow quick and easy queries of phenotypic data and available biospecimens, based on specific research questions.

We have developed a centralized modular IT tool box (Figure 1) that addresses these needs and requirements. Its modular structure allows for a flexible extension and adaptation to specific settings.

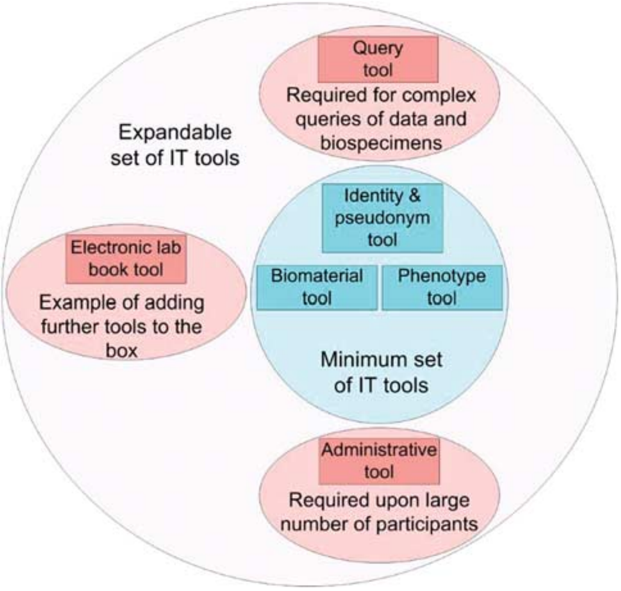

Query tool
Required for complex queries of data and biospecimens
Expandable set of IT tools
Identity & pseudonym tool
Biomaterial tool
Phenotype tool
Minimum set of IT tools
Electronic lab book tool
Example of adding further tools to the box
Administrative tool
Required upon large number of participants

Figure 1 The minimum set of information technology (IT) tools (blue circle) can be expanded by adding other modules (red circles) on demand.

General prerequisites

Informed consent from all study participants as well as harmonized standard operating procedures (SOPs) for all methods and processes are fundamental prerequisites. Storing and freezing capacities need to be in place. If possible, all biospecimens should be stored at one location to assure comparable storage conditions and to minimize cost. For the identification of samples, a barcoding approach (ideally two-dimensional) is preferred. Documentation forms including the sample IDs should be available for printing so as to quickly note any changes concerning the withdrawal or processing of biospecimens if necessary. Internet access is mandatory in order to enter data into the IT toolbox.

IT toolbox

We propose a flexible and centralized IT toolbox as opposed to federated, distributed, or hybrid IT systems as described elsewhere.[11–13] A centralized approach has several advantages: Sites with no expertise and manpower can still participate in large studies, services of data management and infrastructure can be provided by experts. Documented metadata and coding of scales will be standardized, which saves time for mapping and querying and at the same time leads to high-quality data. Centralized databases save cost in maintenance and work. Already existing local IT systems for the management of clinical routine tasks might also be integrated into the proposed centralized IT toolbox; this, however, would require additional mapping steps due to the heterogeneity of metadata.

The IT toolbox, as shown in Figure 1, comprises five tools: (1) an identity and pseudonym tool, (2) an administrative tool, (3) a biomaterial tool, (4) a phenotype tool, and (5) a query tool. A precursor of the toolbox introduced here was described before.[14]

(1) The identity and pseudonym tool stores all identifying data of a participant, including name and contact details, and administers the pseudonyms used to store biomaterial and phenotype data of all participants. To enhance data protection, one participant receives one pseudonym per IT tool during the course of a single study.
(2) The administrative tool is tightly linked to the identity and pseudonym tool. This is essential as the administrative tool should allow for the management of identifying data and of follow-up dates for local participants. It should enable the presentation of family relations needed for family-based studies, and indicate the status of agreement of the informed consent signed by the participant.

(3) The purpose of the biomaterial tool is to manage the collected biomaterials. Data stored within this tool describe the location, type, and quality of each single biospecimen linked to the unique sample ID.
(4) The phenotype tool is used to collect clinical, demographic, and other phenotype data of all study participants.
(5) The query tool enables regular stocktaking of biomaterial, facilitates scientific analyses, allows for the request of biological samples and data as well as for the generation of new research hypotheses.

Needs meet IT toolbox

In order to increase recruitment numbers for psychiatric genetic studies, our aim is to integrate many research sites by providing an administrative framework across different organizations. To this end, it should be mentioned that our approach is not primarily conceptualized as a meta-database[13] but rather as a solution for data collection, management, and analysis of *a-priori* planned scientific projects. Non-academic sites providing only primary care can be integrated into these collaborative research projects with minimal on-site administrative burden. As outlined below, the concept addresses multiple IT challenges:

(a) The need to enable easy participation of new sites at any point of the study is challenged by the complications of a multicentric approach with its different geographical locations and varying IT resources. This challenge can be met by one site which may not even be part of the collaboration and which serves as IT provider for all collaborators.[15] Hence, all five IT tools need to be web-based so that all partners may participate regardless of their geographical location.

(b) In a multicentric research collaboration, researchers may want to analyze and publish their data prior to a joint publication of the collaboration. This should be dealt with through data management policies. From an IT point of view, a very flexible rights and roles concept is the solution. Such a concept can guarantee users to only see and work with their own data before sharing them with the collaborators. A more detailed explanation of the rights and roles concept in the phenotype tool can be found in Supplementary Figure 1. Rights and roles concepts implemented in the proposed IT tools are constructed in such a way that different roles can be defined. Typically, the following roles can be distinguished: administrator, monitor, clinical investigator involved in the phenotypic assessment of the patient, biostatistician, or data manager. On the basis of each single item or data set, different rights in great detail can be assigned to each role, for example, read only, create, alter, allow export, or delete, according to the needs and regulations. Reports, exports, and statistics can also be managed by the rights and roles concept. The concept is applied per IT tool, where each user is assigned a distinct role:

- The administrator will be able to see all data (except for the identifying ones) but cannot change them.
- The clinical investigator will see all data of his patients including the identifying data (from his work package/sub project) and can enter and alter them.
- The biostatistician will be able to read and export all data (except for the identifying ones) but not to alter them.
- A monitor reviewing the data of a clinical trial for quality control purposes will see all data of all work packages except for the identifying ones.

Usually, only the administrator of the respective IT tool is able to assign rights and roles to users. If necessary, this can be more than one role per user.

(c) The requirement of enabling longitudinal assessment of phenotypes and biobanking brings up the issue of correctly mapping pseudonyms. The two main IT tools to solve this problem are the identity and pseudonym tool and the administrative tool. The first checks the identifying data of participants against an existing database and automatically maps all pseudonyms of a participant in one list. The administrative tool simplifies creating pseudonyms for one participant for the different IT tools by just entering his identifying data once (Supplementary Figure 2). Furthermore, the administrative tool reminds the study manager to invite participants to follow-up appointments and manage them.

(d) Although the use of pseudonyms and the separate storage of identifying data can be considered a standard procedure, it may not be sufficient to fend off the risk of identification. This could be further minimized by using different pseudonyms in the phenotype and the biomaterial tools. The mapping table of these pseudonyms should be kept with an honest broker outside the collaboration. This honest broker provides the identity and pseudonym tool as a service. In our experience, institutional review boards favor the idea of a physical and organizational separation of data. This separation and storage of data through an honest broker not involved in the project is also put forward in a generic data protection concept developed in collaboration with German data protection officials.[16] The honest broker can by no means gain access to data other than the identifying ones. The hardware and software requirements for the protection of the identifying data are as follows: The list of the identifying data and the pseudonym mapping list are kept in an enterprise grade database on hardened Linux servers behind a two-staged firewall. Communication between the database and the application servers is strongly encrypted with HTTPS/SSL. Furthermore, the database servers are only accessible from defined IP addresses. Moreover, all IT tools are secured by a login name and a password, which must meet a predefined password strength and is changed in regular intervals, following an automated reminder to do so. Before external researchers can query data with the query tool, all data will be pseudonymized for a second time (Supplementary Figure 3). More importantly, the amount of readily visible information will depend on the level of data accessibility. For example, instead of showing a sequenced genome of an individual participant, a box would be ticked stating ‘sequence information is available’.

(e) To allow for the collection of a variety of phenotypes and all biomaterials, the chosen biomaterial and phenotype tools should offer great flexibility and an easy-to-learn function to parameterize data items. To reduce the risk of data incorrectness or incompleteness, free text fields should be used as little as possible and rules should be implemented to ensure that necessary data items, including plausibility checks (Supplementary Figure 4), are filled in. If, for example, a participant ticks the box ‘male’, the question of pregnancy becomes obsolete.

(f) Straightforward identification and retrieval of biomaterial are two of the most important practical issues in large-scale collaborative biobanking efforts. This involves issues like the choice of tubes, type of barcode labels (Supplementary Figure 5), and location of specimen storage.

(g) Finally, the scientific performance of a large-scale multi-site collaborative effort hinges critically on the possibility of executing quick and easy queries on all data and specimens. This is not a problem with the suggested IT solution for *a-priori* planned projects, where everyone uses exactly the same clinical variables and adheres to the same standards for the documentation of biomaterial. However, if such a project is joined with others, the biggest challenge usually is the heterogeneity of specimens and the problem of inhomogeneous documentation. To address this problem, the unified medical language system integrated into the query tool indicates similarity where it exists. The IT query tool should have a user friendly menu with a drag and drop option for constructing queries and the possibility to save queries which are repeated on a regular basis, for example, the number of DNA samples currently available for a specific phenotype.

The modular setup of the toolbox provides the advantage of being able to flexibly add new tools to the IT toolbox. The electronic laboratory notebook is one such tool. It allows the documentation of all materials, equipment, software, and processes involved in the generation of biospecimens and research results. Such precise documentation, that is, the generation of metadata, is required for long-term preservation of research results.

Flexible toolbox approach

Our design of the IT toolbox is ideally suited to fit the needs of a variety of research organizations. The approach is based on one simple principle: you only take what you need. To illustrate the modular approach, we consider four different cases:

(i) Small monocentric research project, which has not yet collected any data or samples.
(ii) Small monocentric research project, which has already collected a range of phenomic data and biomaterial.
(iii) Large-scale monocentric research project aimed at recruiting a large number of participants; may already have collected a range of phenomic data and biomaterial.
(iv) Large-scale multicentric research project, which has collected and will continue to collect a host of phenomic data and diverse biomaterial. This includes longitudinal cohorts and epidemiological samples in large catchment areas.

For use case I, the minimal set of tools comprising the identity and pseudonym tool, the phenotype tool and the biomaterial tool will satisfy all needs. Participants can be registered, their identifying data securely stored, and all phenomic data and biomaterial can be managed.

In use case II, when the research project has already collected sufficient data for analyses, a query tool should be added to the minimal set of tools.

Most relevant in use case III is the large number of participants to be recruited. This large number makes an administrative tool for organizing follow-up dates, administering the status of the informed consents, and displaying family relations necessary. Moreover, receiving pseudonyms for one participant for all IT data collection tools by entering identifying data only once minimizes the work load.

For the highest expansion stage of a research project (use case IV), the complete set of web-based IT tools should be implemented. Especially the query tool should be on a very advanced level allowing not only for requests for biomaterial and phenomic data from within the research project, but also from external persons and institutions, always in accordance with data protection requirements.

Use of commercially available vs home-made software

In any project that involves a modular use of IT components, one is faced with the question whether to use existing software tools or to develop them oneself. We recommend using professional software, wherever possible, as this allows for continued support, access to updates and bug fixes, and thus guarantees a high level of sustainability. Where no such software exists, one should develop new tools using existing frameworks and technologies. This, in combination with an adaptation of the identity and pseudonym tool, is what we use for the administrative tool. For the phenotype tool, we use secuTrial from interactive Systems (Berlin, Germany), for the biomaterial tool we use Starlims from Abbott (Abbott Park, IL, USA), for the query tool we use i2b2 developed by the National Center for Biomedical Computing in Boston (MA, USA), and for the identity and pseudonym tool we use the PID (Patient Identifier) Generator[17] developed by the German umbrella organization TMF (Technology, Methods, and Infrastructure for Networked Medical Research). The choice of type of an electronic lab book depends on the research focus of the project.

Outlook

The IT toolbox is meant to guide researchers in setting up an IT framework for managing sensitive personal data and biomaterial for psychiatric genetic research. It particularly aims at research collaborations that have received funding for a specific project but have not yet gathered experience in setting up the necessary IT environment. With the toolbox, we present an easy-to-follow guideline on the IT components essential to the management of sensitive data and biomaterial in collaborative longitudinal psychiatric genetic research. A detailed sequential checklist on the implementation of the toolbox essentials is provided (Supplementary Table 1).

The IT tool box described has been developed as the IT backbone of a multi-center research project on genotype-phenotype relationships and the neurobiology of the longitudinal course of psychosis coordinated by the University of Göttingen (http://www.kfo241.de). Precursors and parts of this tool box, which is continuously adapted to changing scientific and legal requirements, have been used for many years by the German Competence Networks Dementia, Congenital Heart Defects, and Multiple Sclerosis, the Research Network Creutzfeldt-Jakob disease, and another Clinical Research Group at the University of Göttingen, focusing on colorectal cancer. In the near future, the proposed infrastructure will also be used by the Göttingen stem cell biobank of the German Center for Cardiovascular Research and the centralized biobank of the Göttingen University Medical Center.

As discussed, the modular structure allows for an easy expansion of the IT toolbox. Adding a tool for handling sensitive genome-wide genotyping or sequencing data, for instance, could constitute a logical next step. The modular flexibility also allows for the exchange of tools that have become obsolete due to new research strategies or legal requirements. In accordance with the challenges encountered in multi-site academic research, the present approach can be considered a flexible, centralized tool box: Whereas it would be ideal for standardized metadata to be collected and stored centrally, the flexibility of our framework assures that experienced and well integrated local IT systems can be kept at distinct sites while their data can still be queried. For the current and planned implementations of our toolbox approach outlined above, the willingness of investigators to use such centralized system is a prerequisite and the central data storage is mentioned in the informed consent. Importantly, the central administration of data does not mean that individual researchers forfeit control over their own data but retain it via the aforementioned rights and roles concept. Indeed, a partly centralized and partly federated solution has been postulated as combining the advantages of cost-effectiveness and ownership issues inherent in centralized and federate systems, respectively.[11,12,18]

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work is supported by the Deutsche Forschungsgemeinschaft through the Clinical Research Group 241 ‘Genotype-phenotype relationships and neurobiology of the longitudinal course of psychosis’ (http://www.kfo241.de; grant number SCHU 1603/5-1).

References

1 Schulze TG. Genetic research into bipolar disorder: the need for a research framework that integrates sophisticated molecular biology and clinically informed phenotype characterization. *Psychiatr Clin North Am* 2010; **33**: 67–82.

2 Sullivan PF. The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron* 2010; **68**: 182–186.

3 Unutzer J, Choi Y, Cook IA, Oishi S. A web-based data management system to improve care for depression in a multicenter clinical trial. *Psychiatr Serv* 2002; **53**: 671–678.

4 Heiman GA, King RA, Tischfield JA. New Jersey Center for Tourette Syndrome sharing repository: methods and sample description. *BMC Med Genomics* 2008; **1**: 58.

5 Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39**: 1181–1186.

6 Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science* 2004; **305**: 183.

7 Kahn SD. On the future of genomic data. *Science* 2011; **331**: 728–729.

8 Warner V, Wickramaratne P, Weissman MM. The role of fear and anxiety in the familial risk for major depression: a three-generation study. *Psychol Med* 2008; **38**: 1543–1556.

9 Weissman MM, Brown AS, Talati A. Translational epidemiology in psychiatry: linking population to clinical and basic sciences. *Arch Gen Psychiatry* 2011; **68**: 600–608.

10 Almqvist C, Adami HO, Franks PW, Groop L, Ingelsson E, Kere J *et al.* LifeGene–a large prospective population-based study of global relevance. *Eur J Epidemiol* 2011; **26**: 67–77.

11 Thorisson GA, Muilu J, Brookes AJ. Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat Rev Genet* 2009; **10**: 9–18.

12 Webb AJ, Thorisson GA, Brookes AJ. An informatics project and online ″Knowledge Centre″ supporting modern genotype-to-phenotype research. *Hum Mutat* 2011; **32**: 543–550.

13 Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J *et al.* BioMart: a data federation framework for large collaborative projects. *Database (Oxford)* 2011; **2011**: bar038.

14 Dangl A, Demiroglu SY, Gaedcke J, Helbing K, Jo P, Rakebrandt F *et al.* The IT-infrastructure of a biobank for an academic medical center. *Stud Health Technol Inform* 2010; **160**(Pt 2): 1334–1338.

15 Helbing K, Demiroglu SY, Rakebrandt F, Pommerening K, Rienhoff O, Sax U. A data protection scheme for medical research networks.

Review after five years of operation. *Methods Inf Med* 2010; **49**: 601–607.

16 Pommerening K, Sax U, Müller T, Speer R, Ganslandt T, Drepper J *et al.* Integrating eHealth and medical research: The TMF data protection scheme. In: Blobel B, Pharow P, Zvarova J, Lopez D (eds). *eHealth: Combining Health Telematics, Telemedicine, Biomedical Engineering and Bioinformatics to the Edge*. Akademische Verlagsgesellschaft Aka GmbH: Berlin, 2008, pp 5–10.

17 Faldum A, Pommerening K. An optimal code for patient identifiers. *Comput Methods Programs Biomed* 2005; **79**: 81–88.

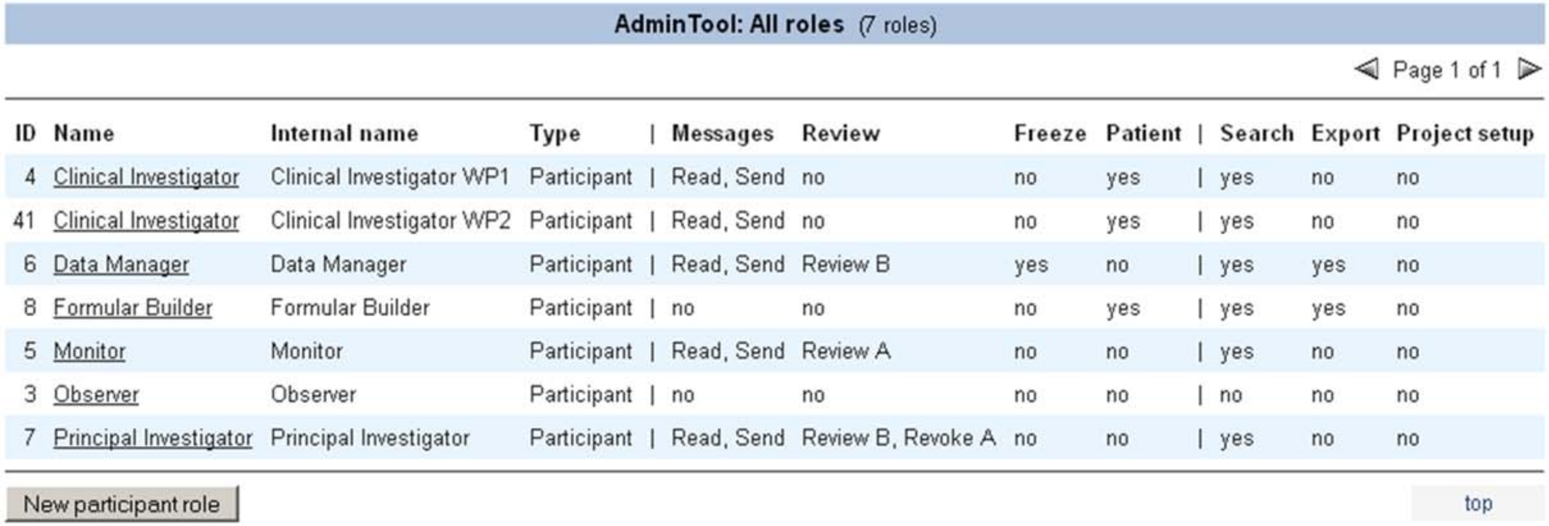
18 Terdiman J, Gul J. PS1-39: The Kaiser Permanente Northern California Oracle Research Database. *Clin Med Res* 2011; **9**: 168–169.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)

IN PRESS WITH MOLECULAR PSYCHIATRY

Supplemental Figure 1. Rights and roles concept as implemented in the software secuTrial (www.secutrial.com). (A) Typical roles in a collaborative psychiatric genetic research project. Note that there are two clinical investigators, one for work package (WP) 1 and the other for WP 2. Both will only see the data of their individual WP. The "messages" option controls whether a role is allowed to read internal system messages or to also send own messages. The "review" rights are essential for validating the data in the database. These rights can be set dynamically according to the project configuration. This is one possible scenario: A clinical investigator locks one or all forms of one visit for further manipulation and sets them to status review A, because from his point of view data entry is complete. The principal investigator checks the entered data, revokes review A, changes several values and finally locks it again with setting it to status review B. The next step could be a monitor who revokes status review B by opening a query on individual questions. Queries can be solved by authorized persons who have the according rights. By this workflow of checks the data quality gets increased until a point where it is decided that data entry is complete. The "patient" option is meant to create new study subjects within the IT tool and add them to a project. The search right enables the role to search items via search forms. With the export right, the whole project data can be exported into a format which is convenient for data analysis (e.g. SAS or CSV). No new project can be set up by any of the roles indicated here; to do this is the responsibility of the IT administrator. (B) Detailed rights of the Clinical Investigator from WP 1. Specific rights can be given and withdrawn. In our example, the SCID interview can be edited and read but not commented.

A Global view on different roles defined in a phenotype tool

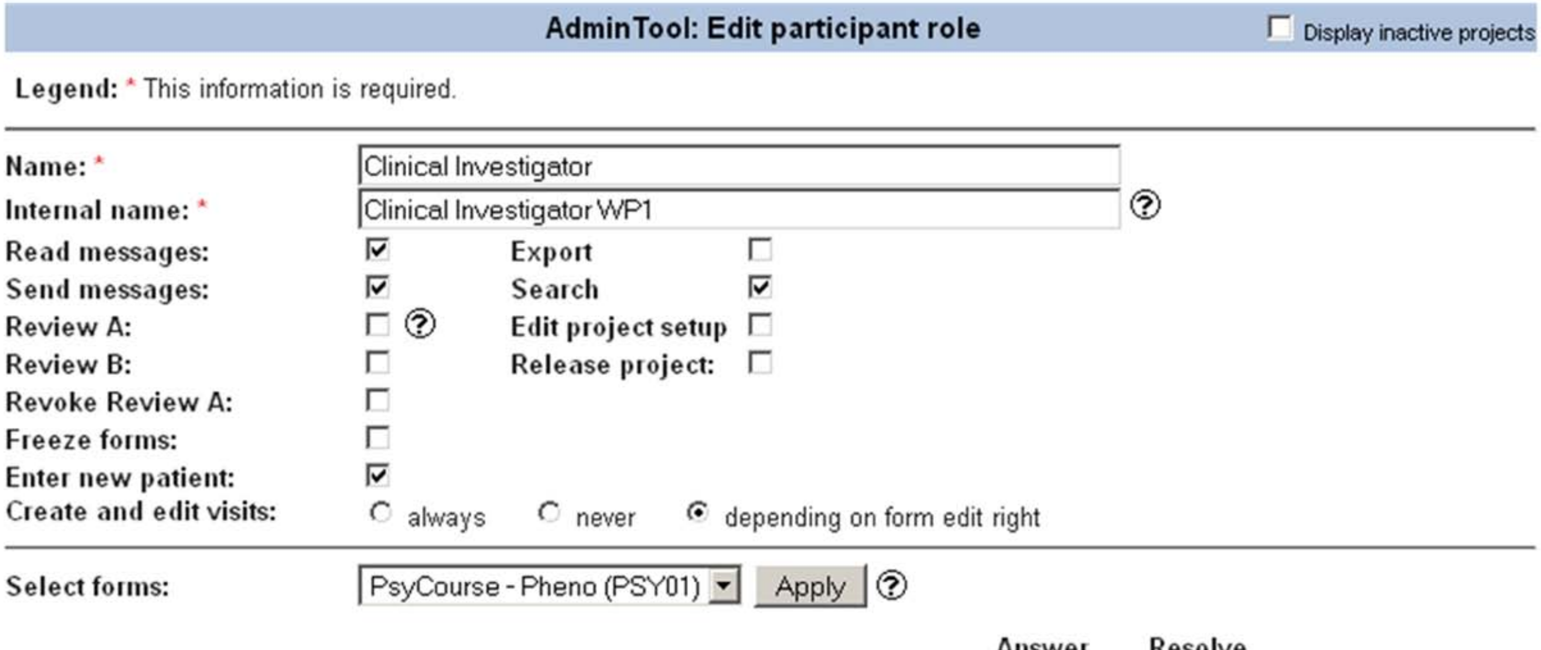


AdminTool: All roles (7 roles)

◁ Page 1 of 1 ▷

ID	Name	Internal name	Type	Messages	Review	Freeze	Patient	Search	Export	Project setup
4	Clinical Investigator	Clinical Investigator WP1	Participant	Read, Send	no	no	yes	yes	no	no
41	Clinical Investigator	Clinical Investigator WP2	Participant	Read, Send	no	no	yes	yes	no	no
6	Data Manager	Data Manager	Participant	Read, Send	Review B	yes	no	yes	yes	no
8	Formular Builder	Formular Builder	Participant	no	no	no	yes	yes	yes	no
5	Monitor	Monitor	Participant	Read, Send	Review A	no	no	yes	no	no
3	Observer	Observer	Participant	no	no	no	no	no	no	no
7	Principal Investigator	Principal Investigator	Participant	Read, Send	Review B, Revoke A	no	no	yes	no	no

B Detailed description of the Clinical Investigator WP1 role



AdminTool: Edit participant role □ Display inactive projects

Legend: * This information is required.

Name: * Clinical Investigator

Internal name: * Clinical Investigator WP1

Read messages:	☑	Export	□
Send messages:	☑	Search	☑
Review A:	□	Edit project setup	□
Review B:	□	Release project:	□
Revoke Review A:	□		
Freeze forms:	□		
Enter new patient:	☑		

Create and edit visits: ○ always ○ never ● depending on form edit right

Select forms: PsyCourse - Pheno (PSY01)

Supplemental Figure 2. Flow of pseudonym mapping and retrieval. (A) Two different pseudonyms are generated for John Doe in the identity and pseudonym tool.

(1) Identifying data of the study participant are entered into the administrative tool.

(2) The identifying data are transferred to the identity and pseudonym tool.

(3) Two pseudonyms for the phenotype tool and the biomaterial tool are generated.

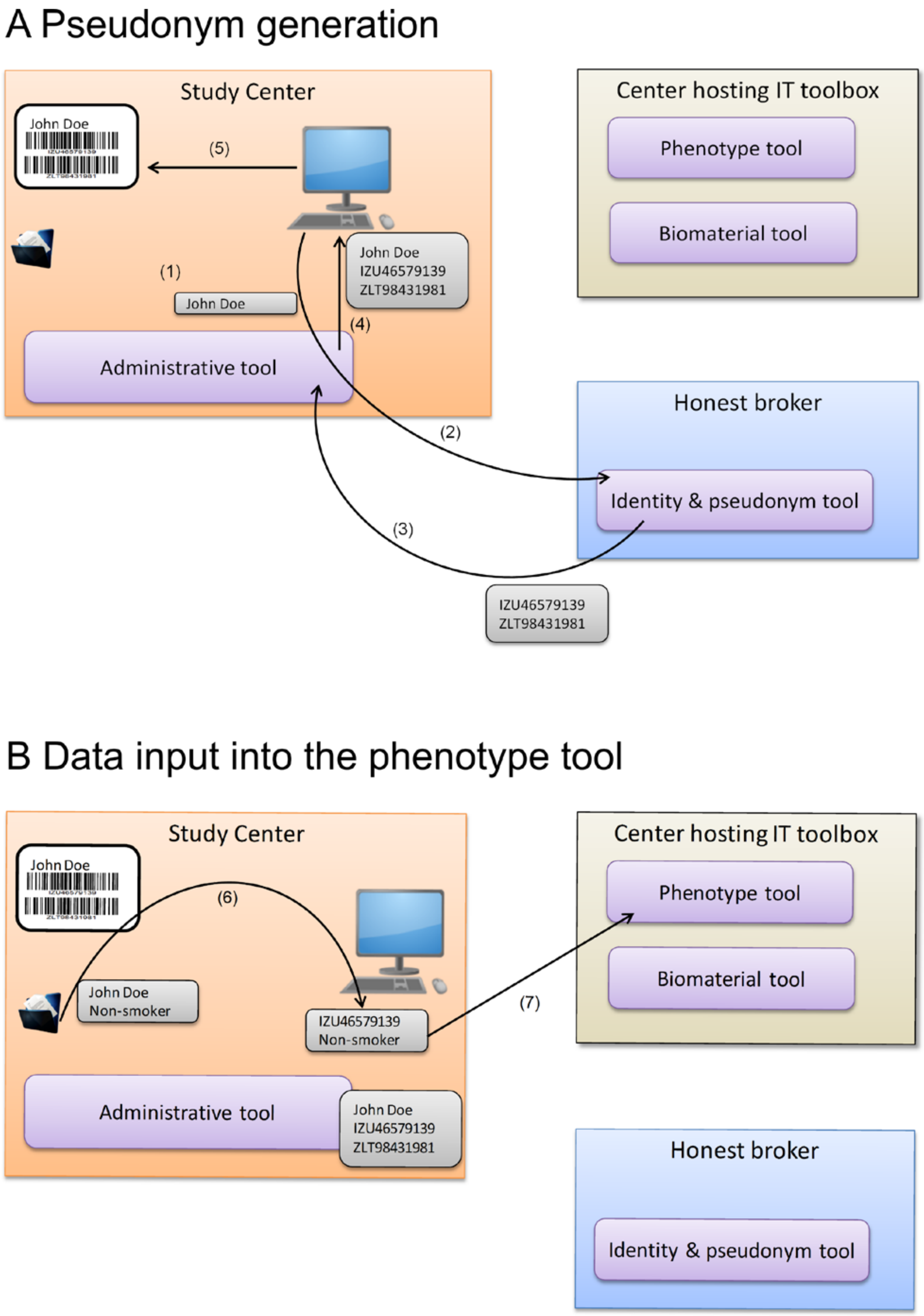
(4) Pseudonyms are linked with the identifying data.

(5) Pseudonyms are printed as barcodes together with the identifying data.

(B) Data input into the phenotype tool is exemplified.

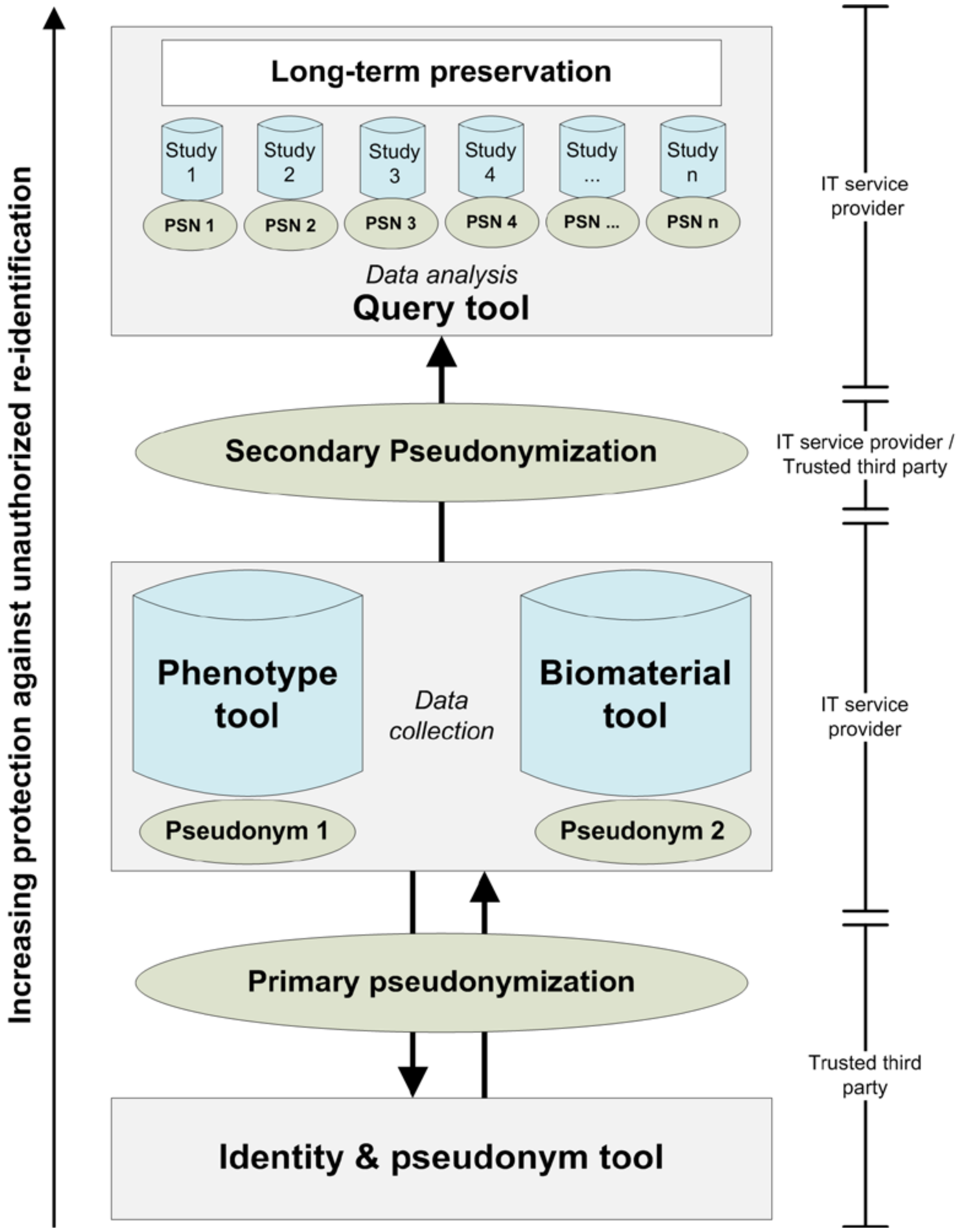
(6) Barcode of the phenotype tool is linked with the phenotype item „non-smoker“.

(7) Pseudonym and the phenotype data are stored together in the phenotype tool.



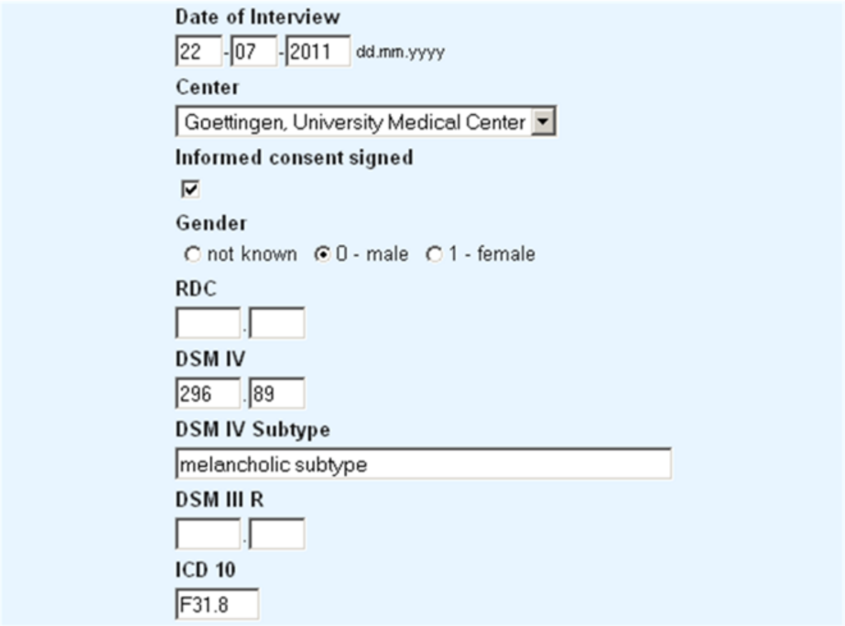
A Pseudonym generation
Study Center
John Doe
IZU46579139
ZLT98431981
(5)
(1)
John Doe
John Doe
IZU46579139
ZLT98431981
(4)
Administrative tool
(2)
(3)
Center hosting IT toolbox
Phenotype tool
Biomaterial tool
Honest broker
Identity & pseudonym tool
IZU46579139
ZLT98431981
B Data input into the phenotype tool
Study Center
John Doe
IZU46579139
ZLT98431981
(6)
John Doe
Non-smoker
IZU46579139
Non-smoker
(7)
Administrative tool
John Doe
IZU46579139
ZLT98431981
Center hosting IT toolbox
Phenotype tool
Biomaterial tool
Honest broker
Identity & pseudonym tool

Supplemental Figure 3. Multi-level pseudonymization. In countries with very strict data protection requirements like Germany, the identity of study participants is ideally kept with a trusted third party. Up to now, this has typically been done through notaries or law offices but is now being replaced by a system in which an academic institution (e.g. a medical informatics department) that is not part of the scientific collaboration acts as a service provider. In such a framework, all data to be collected will be kept in two different IT tools, whereby one participant will have different pseudonyms in both phenotype and biomaterial tool. Before data can be queried, they need to be linked and then to undergo a second pseudonymization step. During this step, the second pseudonym (PSN) is generated and then used in the query tool. Especially for long-term preservation of research data, this is of utmost importance.


Increasing protection against unauthorized re-identification
Long-term preservation
Study 1
Study 2
Study 3
Study 4
Study …
Study n
PSN 1
PSN 2
PSN 3
PSN 4
PSN …
PSN n
Data analysis
Query tool
IT service provider
Secondary Pseudonymization
IT service provider / Trusted third party
Phenotype tool
Data collection
Biomaterial tool
IT service provider
Pseudonym 1
Pseudonym 2
Primary pseudonymization
Trusted third party
Identity & pseudonym tool

Supplemental Figure 4. Front end and back end of the phenotype tool from secuTrial software. In B you can see how the different areas of the questionnaire can be put together.

A Front end of the phenotype tool (end-user view)



Date of Interview

22 - 07 - 2011 dd.mm.yyyy

Center

Goettingen, University Medical Center ▾

Informed consent signed

☑

Gender

○ not known ● 0 - male ○ 1 - female

RDC

.

DSM IV

296 . 89

DSM IV Subtype

melancholic subtype

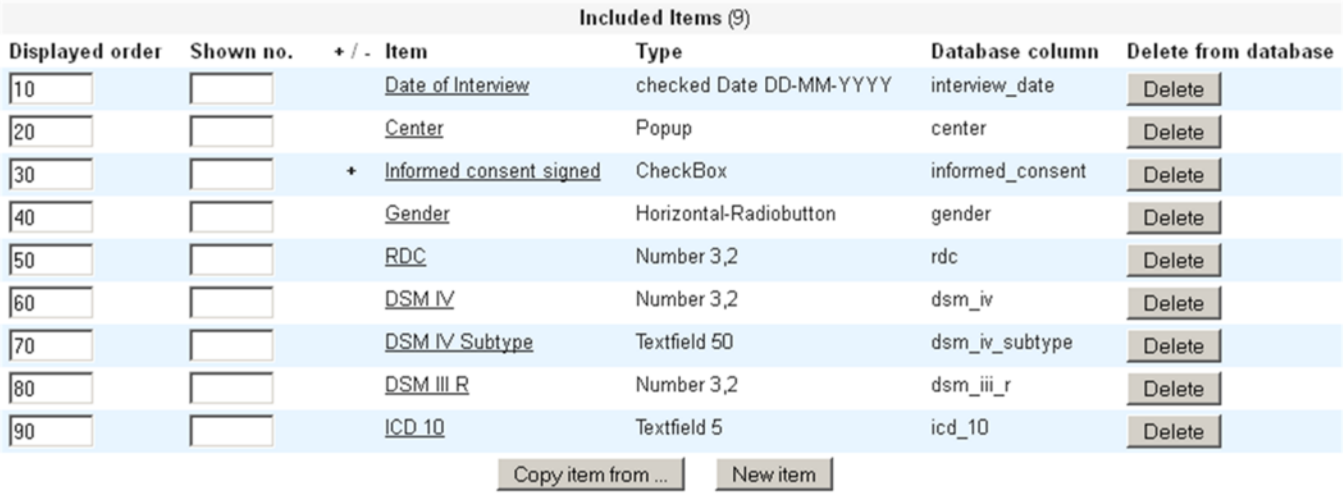
DSM III R

.

ICD 10

F31.8

B Back end of the phenotype tool (IT administrator view)

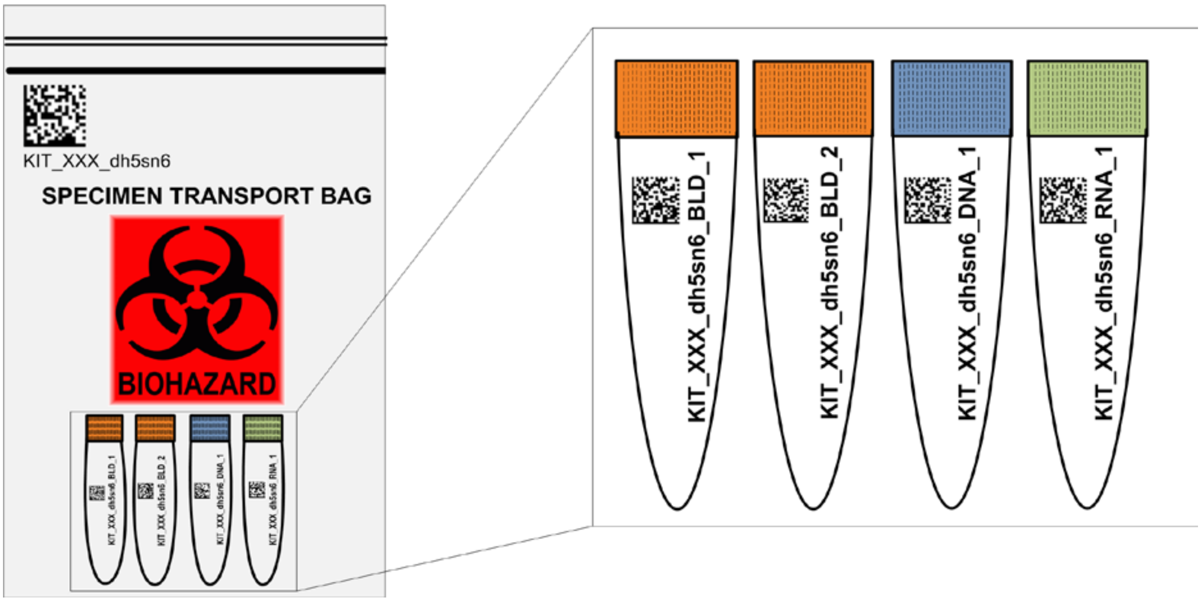


Included Items (9)

Displayed order	Shown no.	+ / -	Item	Type	Database column	Delete from database
10			Date of Interview	checked Date DD-MM-YYYY	interview_date	Delete
20			Center	Popup	center	Delete
30		+	Informed consent signed	CheckBox	informed_consent	Delete
40			Gender	Horizontal-Radiobutton	gender	Delete
50			RDC	Number 3,2	rdc	Delete
60			DSM IV	Number 3,2	dsm_iv	Delete
70			DSM IV Subtype	Textfield 50	dsm_iv_subtype	Delete
80			DSM III R	Number 3,2	dsm_iii_r	Delete
90			ICD 10	Textfield 5	icd_10	Delete

Supplemental Figure 5. Labeling of specimen transport bags and barcoding. The identification of samples is often challenged by unreadable labels, labels that fall off the tubes at low temperatures, or the problem of retrieving samples from a large freezer without knowing exactly where they are. To circumvent these challenges, we recommend using two different 2D barcodes. The first 2D barcode comes with the commercial storage tubes and is permanently laser-etched on the bottom of the tubes. The second one, depicted in the figure above, is created by the biomaterial tool and can be printed onto adhesive labels and stuck onto the side of the tubes. These designed 2D barcodes contain a project ID (XXX), a kit ID (dh5sn6), three letters identifying the type of biomaterial in the tube, based on SPREC (1), and an ascending number. Via the kit ID, all tubes belonging to one participant's visit can be easily identified and stored prior to blood withdrawal in one specimen transport bag which is also labeled with the project ID and the kit ID. 2D barcodes are still machine readable even if about one fourth of the data matrix are lost, ensuring high sample identification rates. To complete the IT infrastructure, a 2D barcode scanner and printer must be available. In addition to 2D barcodes, it is also possible to identify samples by using integrated memory chips. These could be read out wirelessly via radio-frequency identification (RFID) technology. Such memory chips not only store information about the sample itself, like process steps or results of analyses, but also documents and images. With the information permanently stored and physically bound to the sample, the risk of losing essential information is minimized.

1. Betsou F, Lehmann S, Ashton G, Barnes M, Benson EE, Coppola D *et al.* Standard preanalytical coding for biospecimens: defining the sample PREanalytical code. *Cancer Epidemiol Biomarkers Prev* 2010; **19**(4)**:** 1004-1011.


KIT_XXX_dh5sn6
SPECIMEN TRANSPORT BAG
BIOHAZARD
KIT_XXX_dh5sn6_BLD_1
KIT_XXX_dh5sn6_BLD_2
KIT_XXX_dh5sn6_DNA_1
KIT_XXX_dh5sn6_RNA_1

Supplemental Table 1: Sequential checklist on the implementation of an IT toolbox for collaborative psychiatric genetic research

One year before project start:

- Hire IT experts for setting up the IT toolbox
- Describe, document, and compare workflows for recruitment of participants, biomaterial collection, interview procedure, transport of biomaterials and questionnaires, processing and storage of biomaterials, entering of data into databases with existing workflows in the IT tools
- Write a project plan
- Set up a financial plan for the IT toolbox. Software products always have follow-up costs for update and support, hardware, and often additional costs for adaptation
- Write an IT concept
- Buy hardware for the setup of the IT toolbox
- Establish the identification of samples, e.g. 2D barcoded tubes
- Decide on which phenotypic assessment tools (e.g. questionnaires, scales) to implement in the phenotype tool
- Prepare a list of requirements concerning the IT tools and perform a market survey
- Decide on software products for the different IT tools

9 months before project start:

- Write standard operating procedures (SOPs)
- Write patient information and patient informed consent
- Write data protection concept

- Implement the phenotypic assessment tools to be documented
- Buy lab equipment and storage equipment for biomaterial: freezers, liquid nitrogen tanks, centrifuges, tubes
- Instruct facility management to setup N_2 alarm system and power backup for physical biobank room
- Make sure internet access points are available at data entry points
- Prepare a list of required specifications for the adaptation of software tools and make a contract with a company or adapt the IT tools yourself

6 months before project start:

- Approval of patient information, patient informed consent, and data protection concept by the internal review board
- Reproduce the physical storage structure in the tool for the administration of biomaterial and decide on to be documented items for the biomaterial and implement them
- Adapt the administrative tool to the parameters to be documented
- Buy infrastructure-related devices: Barcode scanner, barcode printer, computers to be used at the different data entry points
- Receive changes in the software products and validate the changes
- Decide on a rights and roles concept
- Make a contract with an honest broker for the storage of identifying participant data

Shortly before project start:

- Make sure all plausibility checks and rules in the IT tools have been validated
- Make a backup of the empty IT systems
- Make sure SOPs have all been written and approved

- Ensure that staff is trained in the workflows and in how to work with new machines
- Validate that all IT tools work optimally together
- Distribute logins and passwords to the users and assign roles
- Send empty pre-labeled tubes to all participating partner sites as well as SOPs for external centers
- Train users in the use of IT tools
- Start setting up the query tool
- Start recruiting the first participants

6 months after project start:

- Draw a conclusion about what was good and what went wrong
- Check workflows and implementation of SOPs
- Determine the amount of datasets entered into the IT tools
- Validate the setup of the query tool