TRANSFoRm

Translational Research and Patient Safety in Europe

Deliverable 6.2

Clinical Research Information Model

Work Package:	WP6, WT 6.4
Type of document:	Conceptual framework
Version:	V01
Date:	31 May 2012
Authors:	W. Kuchinke, T. Karakoyun, C. Ohmann, (UDUS)

TRANSFoRm is partially funded by the European Commission - DG INFSO Under the 7th Framework Programme. **(**FP7 247787)
7th Framework Programme http://cordis.europa.eu/fp7/ict/
European Commission http://ec.europa.eu/information_society/index_en.htm

eHealth
Better Healthcare for Europe

European Commission

Index

1. Executive summary

A clinical research information model (CRIM) is presented for the integration of clinical research covering randomized clinical trials (RCT), case-control studies and database searches into the TRANSFoRm application development. TRANSFoRm clinical research is based on primary care data, clinical data and genetic data stored in databases and electronic health records and employs the principle of reusing primary care data, adapting data collection by patient reported outcomes (PRO) and eSource based Case Report Forms. CRIM was developed using the TRANSFoRm clinical use cases of GORD and Diabetes [1]. Our use case driven approach [2] consisted of three levels of modelling drawing heavily on the clinical research workflow of the use cases. Different available information models were evaluated for their usefulness to represent TRANSFoRm clinical research, including for example CTOM of caBIG, Primary Care Research Object Model (PRCOM) [3] of ePCRN and BRIDG of CDISC. The PCROM model turned out to be the most suitable and it was possible to extend and modify this model with only 12 new information objects, 3 episode of care related objects and 2 areas to satisfy all requirements of the TRANSFoRm research use cases. Now the information model covers GCP compliant research, as well as case control studies and database search studies, including the interaction between patient and GP (family doctor) during patient consultation, appointment, screening, patient recruitment and adverse event reporting.

The extension of PCROM was achieved by introducing two high-ranking concepts (areas) into the information model: a care and an ENTRY area. Because some of the research operations take place in the overlapping area of care and non-care activities (e.g. patient consultation and recruitment), a new area for care-related research activities was introduced including episode of care and encounter as basic elements. In the ENTRY area different aspects of data collection are addressed and combined, like the data semantics for observations, assessment activities, intervention activities and patient reporting.

Because the querying of databases plays an important part in the TRANSFoRm research infrastructure and is a prerequisite of patient recruitment and case-control studies, the information model was extended to cover the information requirements for the development of the query tool. The eligibility concept of the model was expanded by an integrated Eligibility Criteria Model (ECM) that specifies the information constraints on the formulation of inclusion / exclusion criteria. A novel combination of Filter and Comparator was introduced for the design of queries for inclusion and exclusion criteria enabling the expression of complex temporal relationships.

The developed CRIM informs the TRANSFoRm provenance, security and interoperability frameworks. It will directly impact the Reference Terminologies and the development of study protocols. In conjunction with the Clinical Data Integration Model (CDIM), CRIM will be used within the TRANSFoRm workbench, for the Query tool and the eCRF tool.

2. Introduction

The TRANSFoRm project is developing a distributed international user-centred platform for the integration of primary care clinical and research activities to enable efficient research with health care data. The basis of the information modelling is two medical use cases, whose workflow was described as sub-use cases with a clear focus on the clinician's point of view. This made a detailed change of the workflow descriptions necessary to present a feasible model. In addition, to achieve an extensible information model, TRANSFoRm clinical research use cases have to be fitted into a general clinical research process model covering GCP compliant randomized clinical trials (RCT) and enable the electronic health record (EHR)-based clinical research platform to adapt to the complex needs of clinical research use cases with data collection from eCRF, EHR and web questionnaires, as well as data collection from primary care databases. In TRANSFoRm a clinical research information model (CRIM) is needed for the integration of the information flow in clinical research, including randomised clinical trials with the reuse of primary care data and the use of eCRF data (according to eSource Data Interchange) [4], with TRANSFoRm's data model, integration model, provenance framework and security framework. For this purpose, different relevant standards on clinical research information representation, such as BRIDG, HL7 RIM, CTOM (CTODS), STDM, SDM, CDASH and openEHR (see later in document), were used to identify their applicability for the information model.

3. Methods and approach

Three levels of modelling

Our approach consists of applying three levels of modelling and using Use Case driven Object Modelling with UML [2]. The basis for the information model is the two TRANSFoRm medical Use Cases Gastro-Oesophageal Reflux Disease (GORD) and Diabetes (DT2) described in detail in WT 1.3 [1]. The workflow of both use cases were described in UML activity diagrams. The single sub-use cases were modelled separately using MS Visio™ and Enterprise Architect™. In a next step using Enterprise Architect™ the sub-use cases were combined and streamlined to cover the entire research process. Workflow modelling is an established technique for business process re-engineering. In the clinical research domain the use of workflow modelling is still used only rarely [5]. But often clinical trial modelling in UML can serve as a standard format because it allows the detailed description of research processes in Activity Diagrams to enable a process analysis [6]. On the other hand, the Business Process Modelling Notation (BPMN) [7] is used increasingly to model business procedures in a graphical notation that can facilitate the understanding of collaborations and business transactions. We decided to use UML for the use case workflow modelling and not BPMN or a standardised and computable workflow representation, because our information model has only to communicate ideas and the tools

using the model are only model-based. In addition, BPMN lacks some expressivity and flexibility that is needed in our model and PCROM is already UML based.

The information model has to adapt to the requirements of the two very different TRANSFoRm research use cases spanning a broad area of medical research. They cover the use of existing routine healthcare data for research purposes. But, whereas GORD is designed as a Randomized Clinical Trial (RCT), Type 2 Diabetes is a non-interventional study based on existing data using a case-control design. Both, the developed information model and eligibility model are represented in UML class diagrams which were created with Enterprise Architect™.

It is the basic idea of our workflow-based approach that it is in general necessary to consider the clinical research workflow in depth in the development of information models, because the reuse of care data for clinical research results in complex relationships in the activity flows including branching, merging and linking. An information model in software engineering is a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse. It can provide sharable, stable, and organized structure of information requirements for the domain context [8]. In this way, CRIM will provide the necessary information requirements for the tools developed in TRANSFoRm.

A three layers approach was used for modelling: UML based use case driven modelling was used first, for a static line of modelling with diagrams describing concepts in the clinical research domain-of-interest, and second for a dynamic line of modelling with activity and sequence diagrams of data collection and the research processes including eSource scenarios.

To build a semantic foundation for the modelling and to guarantee that the information model is compliant with GCP and the RCT process, two diagrams were developed to define the domain-of-interest: first, a simple domain analysis model (DAM) of GCP trials and second, a Reference Clinical Trial Process Model. The diagram for the GCP DAM was created using a simple relationship diagram, and not by using an UML notation. Key terms of GCP [9] were put into relation to each other. The Reference Clinical Trial Process Model was developed as UML activity diagrams. The complete process flow for a GCP compliant RCT was displayed based on the experience with clinical trial conduct at University of Duesseldorf and ECRIN (European Clinical Research Infrastructures Network) and on the requirements described in GCP DAM. Because GCP deals mainly with the protection of trial participants and the quality of clinical trials data, these two areas were identified with relevance for the clinical trial process reference model. Such domain analysis model (DAM) and Reference Model is in general used for conveying an understanding of the domain and allowing that understanding to be assessed and used by others. A shared semantic view is essential if the clinical research community and the healthcare community together with software developers want to achieve computable semantic interoperability (CSI) as basis of interoperability between applications, as is the case in

TRANSFoRm.

Workflow-based approach

There is a movement from a more static representation used for information models to a more dynamic representation that considers quite complicated workflow processes associated with the primary care processes. The primary care domain is so complicated and diverse that here a single information model may not be available [10]. For example, EHR and CRF are not only data input interfaces but they may trigger the generation of reminders, the distribution of invitations, and alerts and they are active according to defined time points on the study timeline (visits and appointments). Or, only access to research data in a database may be allowed, but not the extraction of data from the database. Therefore, it is necessary to consider the clinical workflow more heavily in the development of information models. For construction of the workflow processes UML activity diagrams were used.

Because TRANSFoRm aims for a distributed international user-centred platform for the integration of primary care clinical and research activities to enable efficient research with health care data, an extensible information model is needed. Both TRANSFoRm clinical research use cases have to be fitted into a general clinical research process model to cover GCP compliant RCTs to enable the clinical research platform to adapt to the complex needs of clinical research employing CRF, EHR, web questionnaires and database querying.

These activity diagrams were then used for a “walk through” relating all objects and the processes of the known information models (BRIDG, CTOM, CDISC) and especially to PCROM. The relevance of different assignments was discussed, and it was considered if existing entities and elements are appropriate, in terms of classes in the primary care domain.

Our use case modelling process identified both information required at each stage of the research process and details of the workflow. During this process, it became clear to us that the existing information models insufficiently incorporate primary care research of the workflow requirements into their models. Through the modelling process we developed a complementary set of information objects which have a conceptual and relational connection with the workflow activities at the overlapping area between care and research (e.g. alerts (alarms), appointments, reminders, web questionnaire,…). The concepts of appointment, alert, etc. could also have importance for the information flow for the planned EHR based decision support.

4. Results

GCP trials domain-of-interest

Good Clinical Practice (GCP) is an international quality standard that was developed by the International Conference on Harmonisation (ICH), which governments can transpose into regulations for clinical trials involving human subjects. The diagram for the GCP domain-of-interest was created using a simple relationship diagram (fig. 1). GCP includes a standard on how clinical trials should be conducted, defines roles and responsibilities of clinical trial sponsors, clinical trial investigators, and monitors. Because GCP deals mainly with the protection of trial participants and the quality of clinical trials data, these two areas were identified with relevance for the clinical trial process reference model: Related to patient security is the confidentiality of records that can be used to identify patients, legally acceptable representative, the impartial witness, compensation to subjects and investigators, blinding and randomisation, recruitment, selection and withdrawal of patients, vulnerable subjects, the trial subject in general, well-being of patients, and the medical care of trial subjects. Related to data management are the trial documentation, essential documents, source data, audit trail, CRF, remote electronic trial data systems, security system, and data back-up.

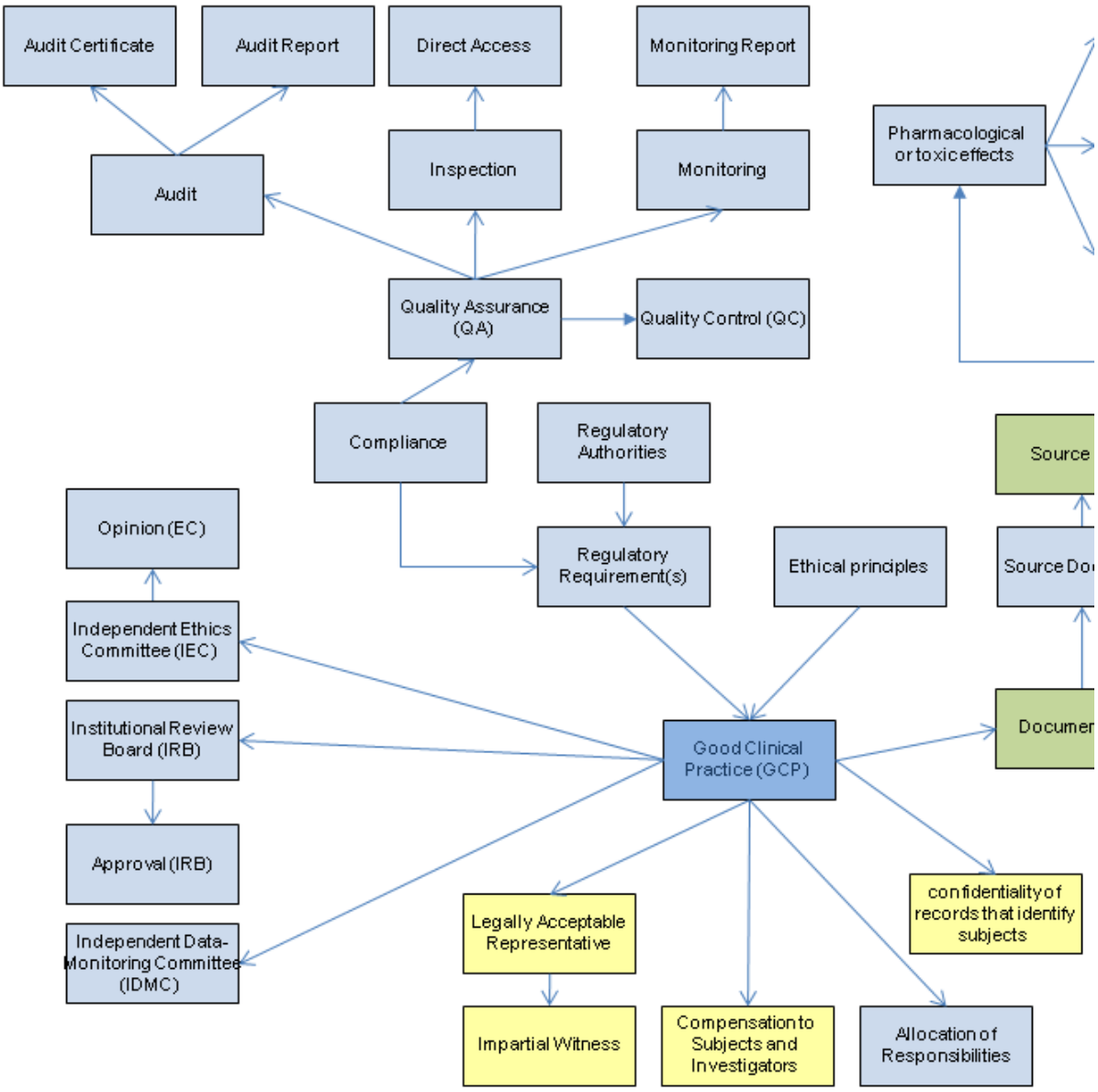
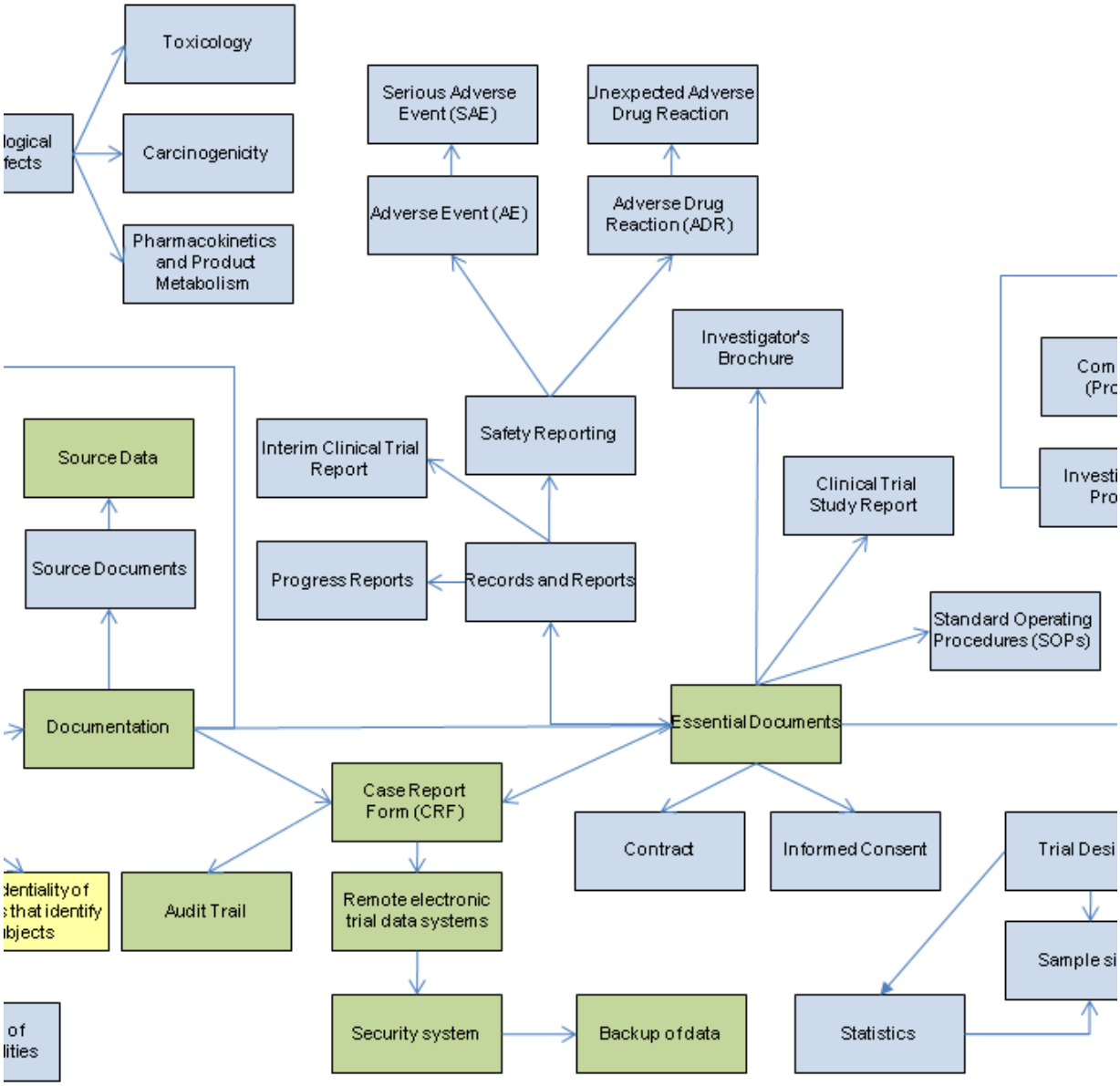
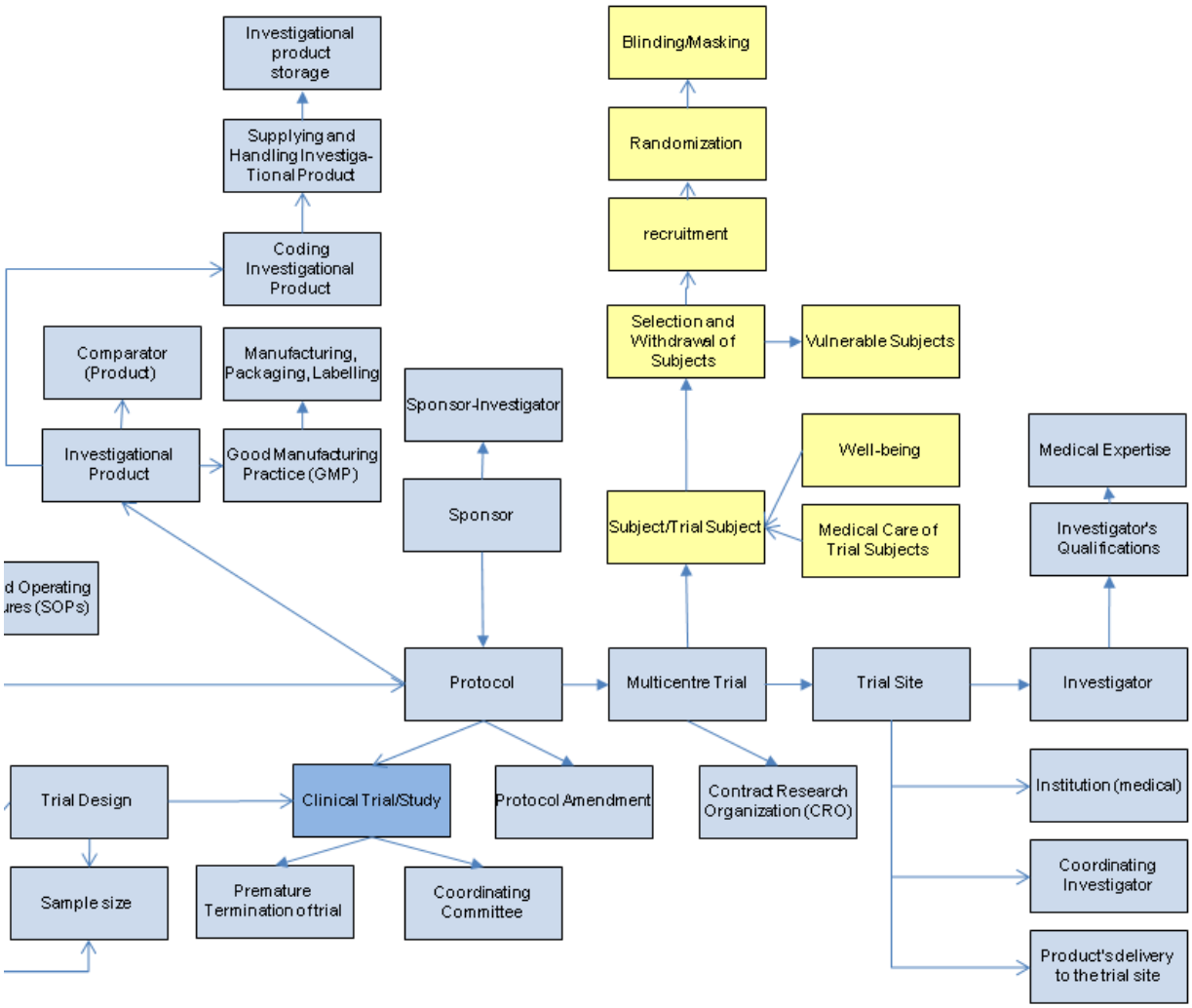


Fig.1 Domain Analysis Model of GCP trials (yellow boxes: patient and trial participant related, green boxes: data management related, grey boxes: other GCP terms)


Toxicology
Carcinogenicity
Pharmacokinetics and Product Metabolism
Serious Adverse Event (SAE)
Unexpected Adverse Drug Reaction
Adverse Event (AE)
Adverse Drug Reaction (ADR)
Investigator's Brochure
Safety Reporting
Interim Clinical Trial Report
Source Data
Source Documents
Progress Reports
Records and Reports
Clinical Trial Study Report
Standard Operating Procedures (SOPs)
Documentation
Essential Documents
Case Report Form (CRF)
Contract
Informed Consent
Audit Trail
Remote electronic trial data systems
Security system
Backup of data
Statistics

(continued)



(continued)

Reference Clinical Trial Process Model

The Reference Clinical Trial Process Model describes all processes that have to be considered to conduct a GCP compliant trial (RCT). From the beginning on in TRANSFoRm, not only clinical studies, but also randomised clinical trials have been considered. This has major consequences: because GORD is a RCT, the GORD use case description is not comprehensive to describe the entire trial. Many components that are decisive, like ethics committee, competent authority, trial master file, monitoring, are not mentioned in the use case description and must therefore

become part of the GORD trial protocol. Based upon the DAM which provided a complete glossary of the GCP standard, a comprehensive delineation of the clinical trial workflow was developed. The GORD sub-use cases were incorporated in the corresponding location of the clinical trials process steps of this clinical trials reference model (fig. 2). As can be seen, the GORD use case covers only the elements of patient identification / recruitment / informed consent and randomisation / CRF / data management. However, CRIM has to consider the entire clinical trial process as described in the reference model. This was achieved, by extending the PCROM model that can easily be mapped to the reference model and to BRIDG (see later).

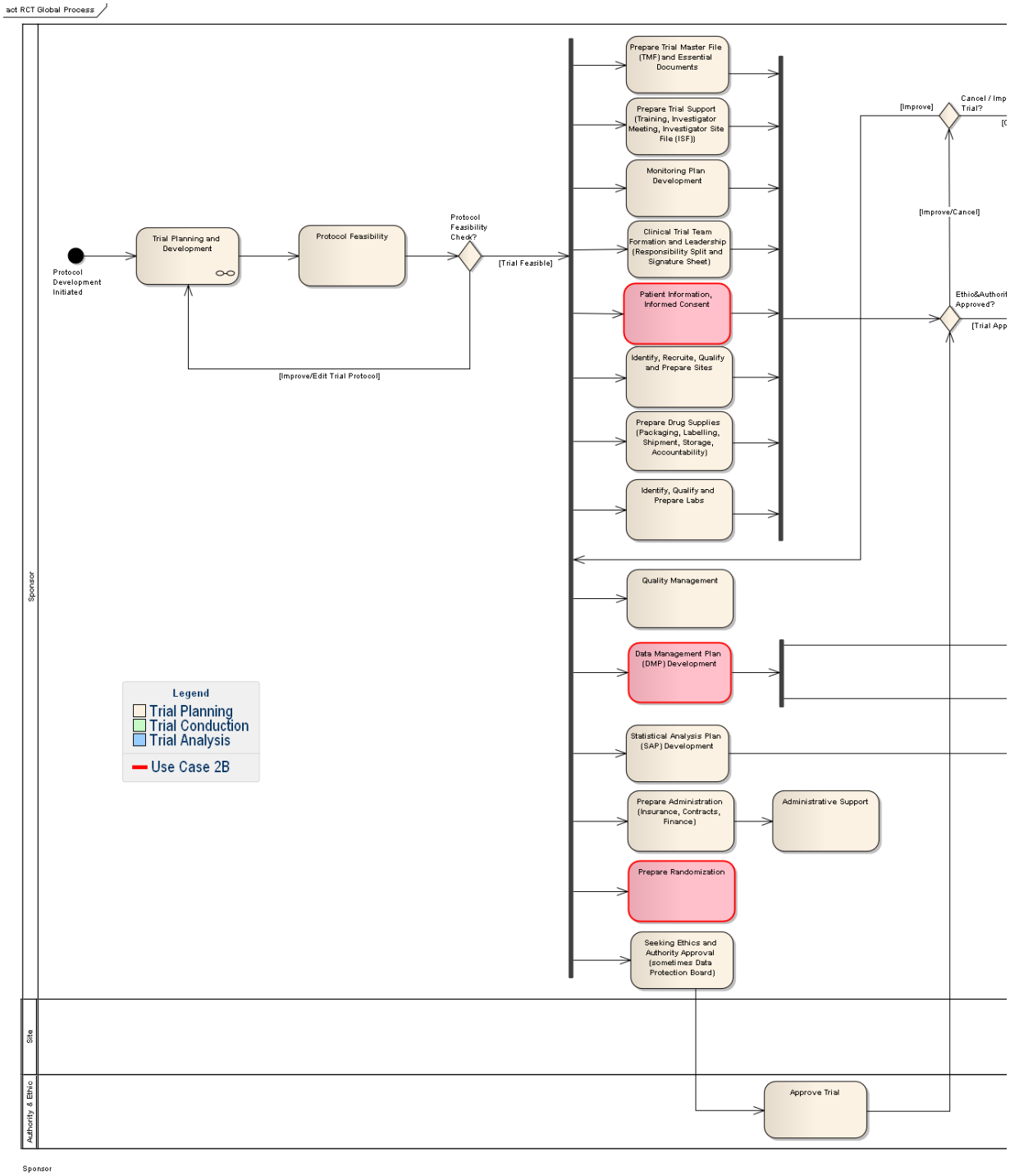
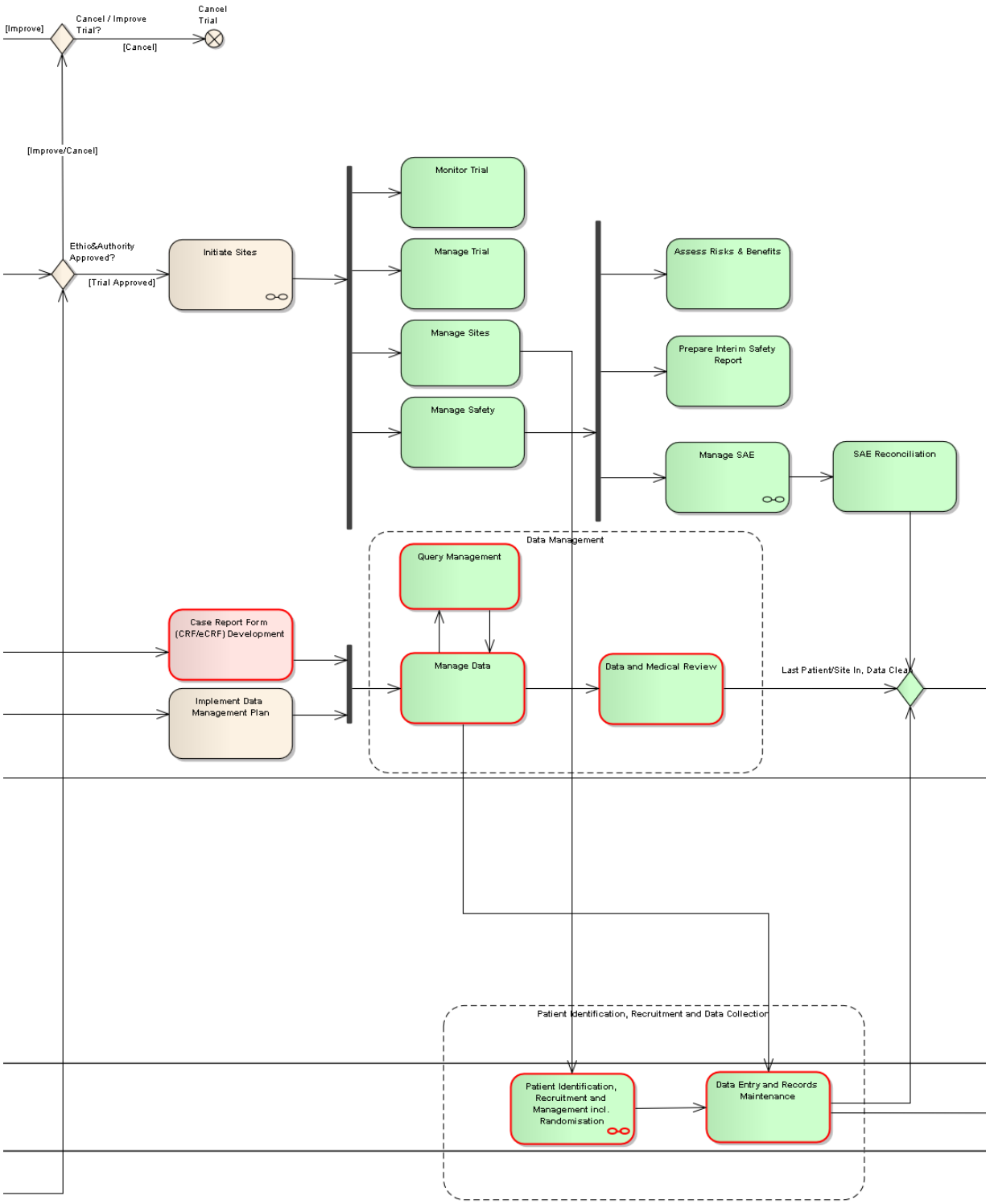
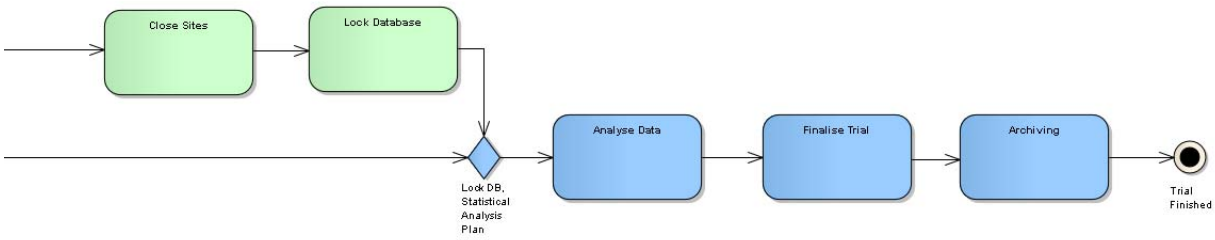
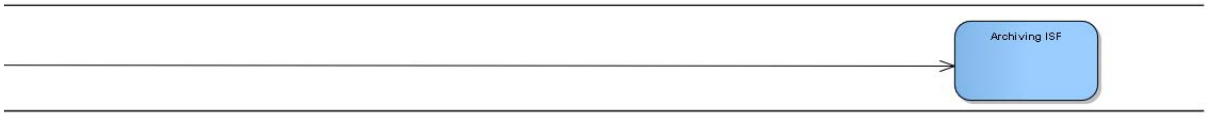


Fig. 2 RCT reference process model (red bordered components are described in the use cases). The clinical trial proceeds from the left to the right (see legend). Areas for data management, patient identification and recruitment that play an important role in TRANSFoRm are indicated.


Cancel / Improve Trial?
[Improve]
[Cancel]
Cancel Trial
[Improve/Cancel]
Ethic&Authority Approved?
[Trial Approved]
Initiate Sites
Monitor Trial
Manage Trial
Manage Sites
Manage Safety
Assess Risks & Benefits
Prepare Interim Safety Report
Manage SAE
SAE Reconciliation
Data Management
Query Management
Case Report Form (CRF/eCRF) Development
Implement Data Management Plan
Manage Data
Data and Medical Review
Last Patient/Site In, Data Clean
Patient Identification, Recruitment and Data Collection
Patient Identification, Recruitment and Management incl. Randomisation
Data Entry and Records Maintenance


Close Sites
Lock Database
Lock DB, Statistical Analysis Plan
Analyse Data
Finalise Trial
Archiving
Trial Finished


Archiving ISF

(continued)

5. The GORD use case

Introduction

GORD represents a RCT study of the effectiveness of a continuous vs. on-demand use of acid suppression on symptoms and quality of life in gastro oesophageal reflux (GORD) with 1 year follow-up (fig. 3) [1]. The aim of this study is to determine the effectiveness in terms of symptom control and QoL and event-initiated assessment of this information in subjects with GORD. An important part of the use case is the identification and recruitment of eligible patients. Eligible patients are predominant GORD cases with both heartburn and acid regurgitation that need PPI treatment, without serious co-morbidity. Study participants will be identified by different methods already before the start of the randomization: firstly, by data mining in EHR on PPI consumers during the previous year, secondly, by identifying patients requesting PPI prescriptions during an appointment (Consecutive PPI prescriptions) and thirdly, by new reflux patients according to inclusion criteria. Before the study begins a screening phase is initiated (PPI washout and PPI “test”). Patients taking PPI will have a washout period including a test use of PPI, H2-blockers and antacid long enough to fulfil the inclusion criteria as reported via web questionnaire. This screening phase is followed by the inclusion of patients into the study by randomisation. Responders will be randomised into two groups: continuous (20 mg Omeprazole/day) or on demand PPI. Outcome will be measured by different ways: symptoms, QoL, alarm symptoms and signs. At entry visit (T= 0) and at month 3, 6, 9 and 12, information about PPI consumption is collected by web questionnaire and/or EHR (structured data collection). In addition, event driven data collection from EHR/eCRF will be performed (e.g. in case of AE).

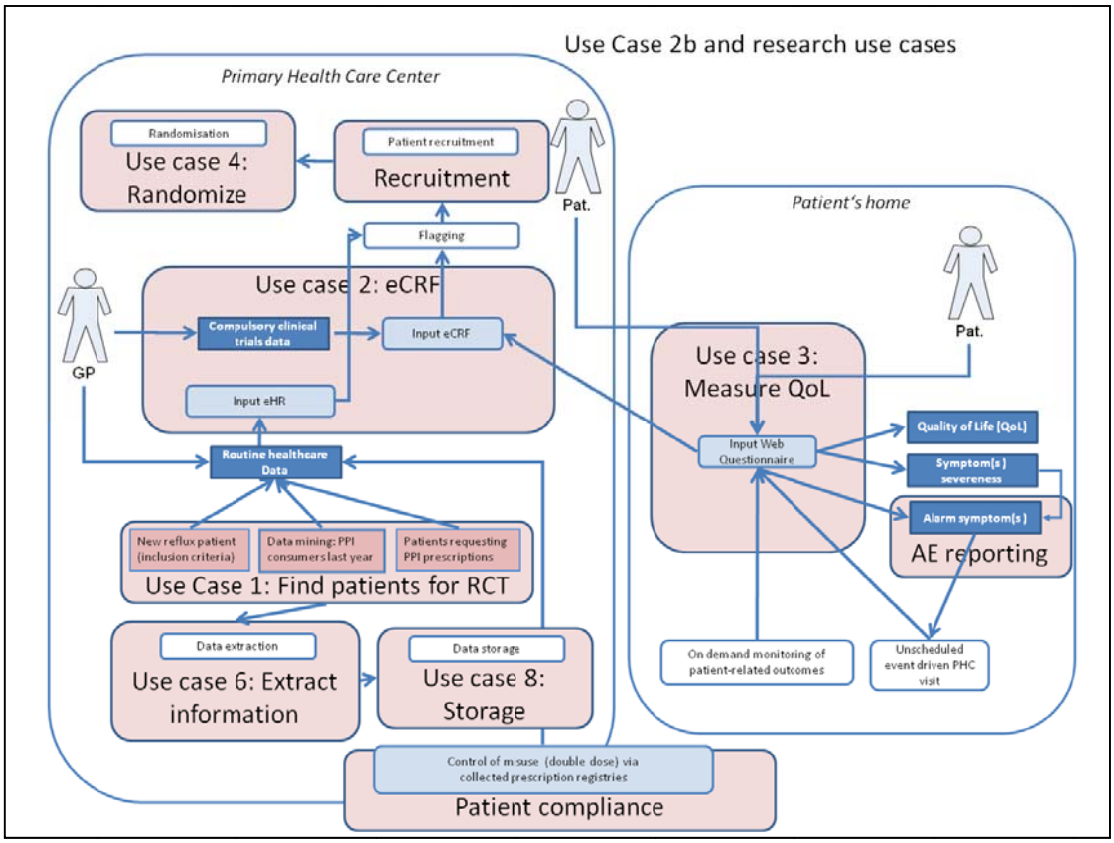


Fig. 3 Relationship of sub-use cases of GORD to the type of data collected, the place of data collection and the reporting of adverse events (Pat. = patient, GP = General Practitioner, QoL=quality of life, brown: sub-use cases and the additional areas "recruitment, patient compliance, AE reporting, necessary for clinical trials, blue: different types of data)

General storyboard

At the Primary health care centre (PHC) or the GP practice the patient is flagged when PPIs are prescribed or a reflux code is entered into the EHR. After an informed consent is given, the eCRF for the study is activated and the impaired QoL due to GORD is assessed through the Web questionnaire. Then, the two week PPI responsiveness period starts. If all inclusion criteria are fulfilled, the patient is enrolled into the study and answers the Web questionnaire at definite time points. The eCRF is filled out at study start (0 months) and study finish (12 months), and whenever there is an event (visit to PHC/ practice) or AE symptom detected. An alert is triggered in case a patient for example consumes double doses continuously. To detect these cases for prescribed PPI, search in registries of collected prescriptions can be used.

A number or roles are described in the use case [1] that had to be adapted to the needs of the workflow (table 1). Especially for data extraction and linking, roles and concepts for the TRANSFoRm privacy framework were introduced. Following roles were used to describe the

GORD workflow: researcher, data controller, TTP, EHR, GP, eCRF, patient web questionnaire.

No	Role in use case description	Description	New role in modeling
1	recipient	• may be an institution/server that receives the requested dataset to which researchers will have access	recipient (institution, database like NIVEL,…), temporal storage place for linked data sets, may provide access to linked data sets, applies a privacy filter to the linked data sets before sending them to researcher
2	primary care database	• data source	data controller (is responsible for data)
3	authority	• authorizes requests from a researcher to extract and link data • consults with the participating data source advisors • presentation of conditions / procedures under which the request will be authorized (e.g. informed consent, permissions,…)	data controller (is responsible for data), and for authorization of access and linkage
4	data source advisor	• responsible for protecting the interests of the data sources; • provides information when specific questions are asked (concerning content, quality, privacy)	necessity of an informed consent is under the responsibility of data controller
5	data sources	• interact with the linker in order to obtain a unique patient number;	data sources reside in different non-care sub-zones; may offer services,

		• interact with the system in order to select patients	like patient selection or data extraction
6	researcher	• accesses the recipient's data, defines the population and the variables which need to be extracted • indicates whether data should be linked at the level of individuals, etc.	researcher
7	linkers	• both national and international (at the international level linking of data at individual level could be performed by an external trusted third party or by the Transform system)	TTP

Table 1: Roles described in the use case and their adaption to the workflow description

The GORD use case is characterised by a screening period before patients are recruited and distributed to one of two different treatment arms. At different time points data collection happens; in addition event triggered data collection takes place. Data collection is done by using eHR, eCRF and a web questionnaire for patient reported outcomes (fig. 4).

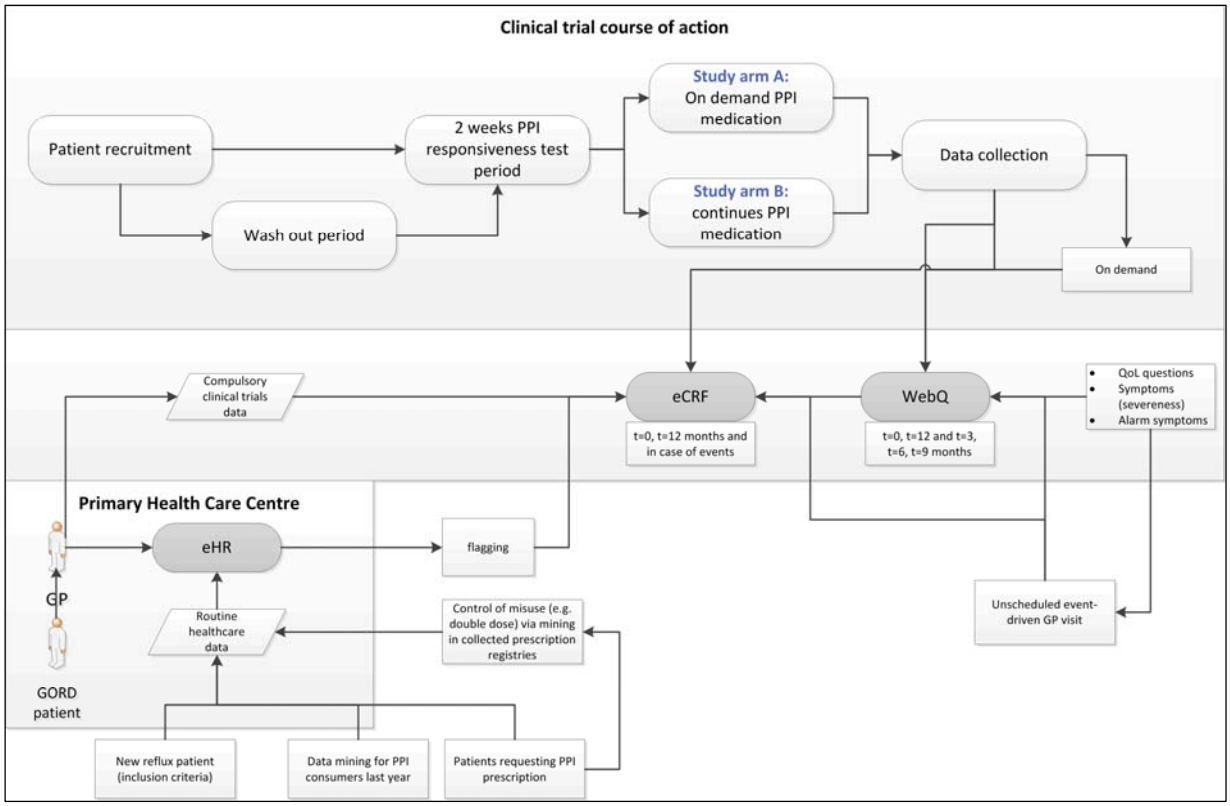


Fig. 4 Relationship between different types of data collection with both treatment arms and data collection time points. WebQ = web questionnaire, t = time point, GP = general practitioner, eCRF = electronic case report form, eHR = electronic health record

The medical GORD use case is divided in a number of sub-use cases for modelling (fig. 3). Central for the information modelling is to consider the processes that deal with the identification of patients, linking a web questionnaire to the eCRF/EHR, the collection of all compulsory data for the study using eCRF, and the linking of EHR data with health care databases. In these areas clinical care processes overlap with clinical research requirements.

Governance

Access to data from care databases and prescription databases are subject to data protection laws and additional policies. It may therefore not be possible that datasets from different data sources can be brought together in a new TRANSFoRm database. Only access to data may be allowed, or data may be stored at a third party (TTP) associated with a database and provided for analysis. In general, researchers can access anonymised data freely, because anonymised data are not subject to data protection laws. All data access, transfer, linkage, encryption and storage require clear policies/protocols, which are different in different European countries.

Concerning the governance issues associated with a clinical trial, GCP and national regulations apply to the conduct of the trial and the collection of data by eCRF and QoL. It must be specified, what data is recorded in the EHR (care area) and if a quality control step is necessary for QoL data input into eCRF. Participation in a trial must be preceded by informed consent.

Use case 1: Find possible participants

Possible participants for RCT can be identified in two ways: by an alert and by a database search (fig. 5). Patient selection may be done either by data mining in primary health care registries or by identifying patients consecutively at the visit at the GP (EHR database). This line of action requires a list of criteria on GORD symptoms identification that specifically fit the study aim, and sufficient data availability in selected care databases. Triggers play an important role. In case of data mining, the search for potential study participants is initiated by the researcher. In case of patient recruitment, the search is initiated by the researcher, and recruitment is initiated when defined symptoms or corresponding codes are entered into the EHR (symptom, diagnostic codes, and prescriptions) or linked databases (self-reported symptoms, prescriptions etc.).

1. The researcher decides upon criteria for study population
2. The researcher initiates data mining or patient tag systems for consecutive recruitment
3. The system selects eligible subjects
4. Patients not fulfilling all further criteria excluded (may be another use case)

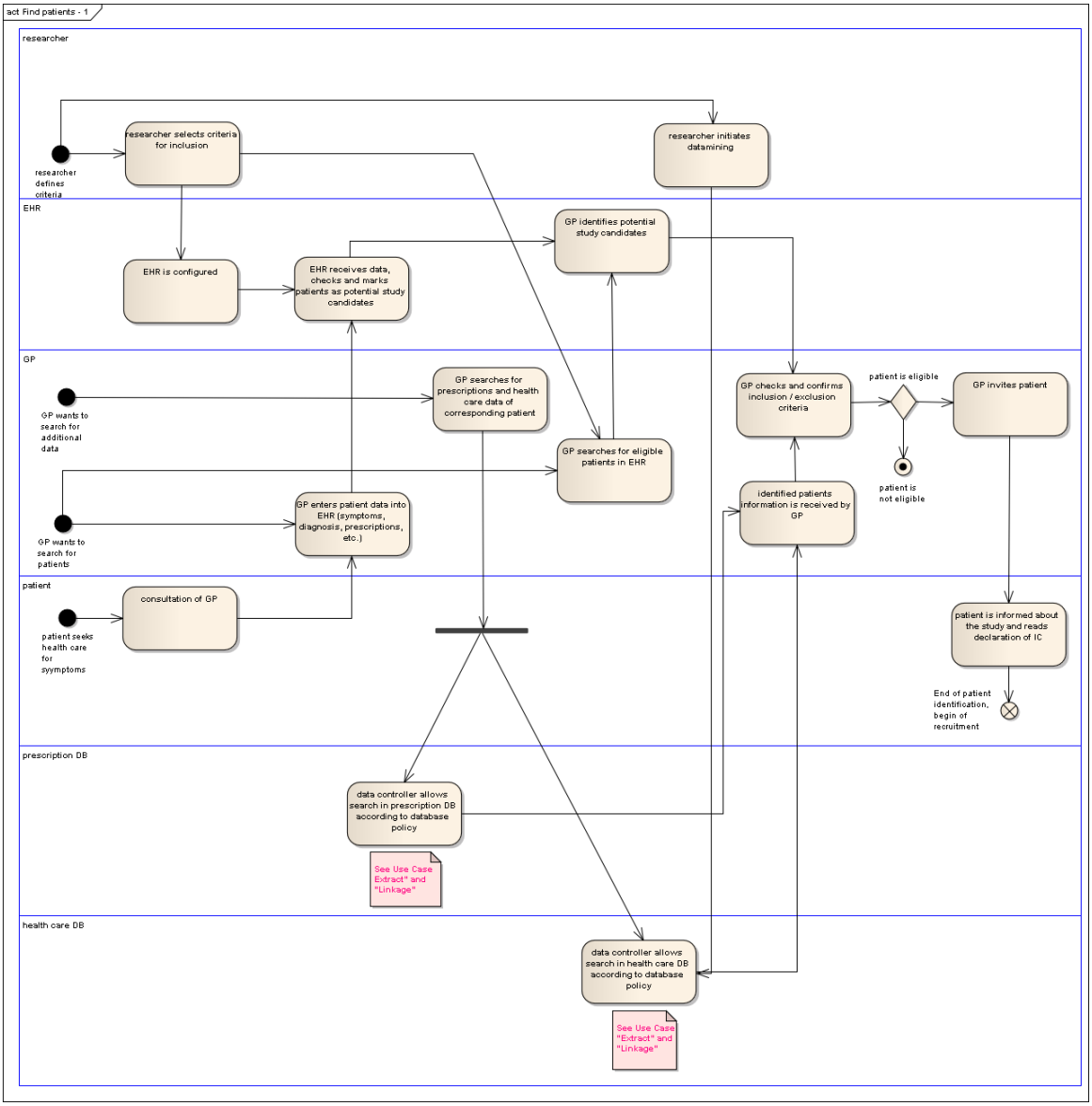


Fig. 5 GORD use case "find patients". Possible participants can be detected by different means (flagging in EHR, search in prescription database,…)

Use case 2: eCRF, capture of important and compulsory data for the study by the GP

For the capture of compulsory data, the eCRF has to be linked to the system used to identify these patients and the eCRF has to interact with the EHR to control that compulsory information is present (eSource scenario).

1. Patient is identified from use case "find patient".
2. eCRF is activated at each PHC visit.
3. GP fills out all compulsory data in eCRF or directly in eHR

In the GORD use case the eCRF plays an important role for data collection of trials data and the integration of ePRO. In this regards it is in accordance with the FDA concept of an integration eCRF hub [11] (fig. 6). The eCRF is used to integrate data from different sources, for example from laboratory and ePRO, in addition to its classical role of collecting clinical study data from the investigator.

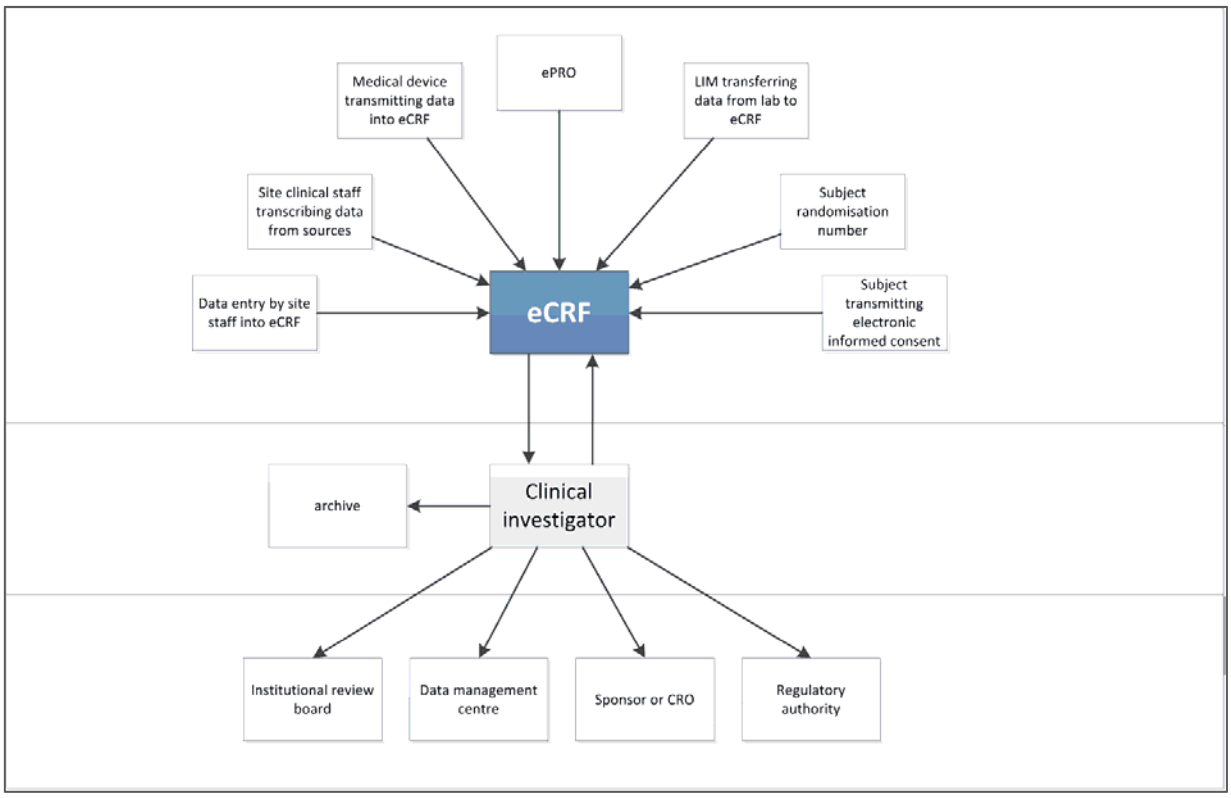


Fig. 6 FDA concept for the eCRF as central information hub (from [11])

This FDA guidance recommends practices that will help ensure that eSource data and source records are accurate, legible, original, attributable (e.g., user name and password), and meet the regulatory requirements for recordkeeping and retention. The following specific topics related to electronic source data are of importance:

1. The identification of the data element as the basic unit of information in the electronic case report form
2. The description of a source of each data element
3. Information about the electronic creation, modification, transmission, and storage of source data and documents;
4. Investigator responsibilities with respect to reviewing and archiving electronic data
5. sponsor responsibilities with respect to reviewing and archiving electronic data
6. Transmission of the data to the sponsor and/or other designated parties

7. Preservation of data integrity

Use case 3: Measurement of Quality of Life (QoL) by PRO (Patient Reported Outcome)

Patient reported QoL, symptoms and their drug consumption is collected with a web questionnaire from the patient' home (fig. 7). In addition the patients fills out the web questionnaire at the first visit (fig. 8). The trigger for the QoL is that a suitable patient has been identified. The patient fills out the web questionnaire and this data is stored. This steps starts after the washout and PPI responsiveness testing phase has been finished and the patient is recruited (informed consent). At the beginning of the trial the web questionnaire is activated for QoL. And the patient answers the web questionnaire at t=0, t=3, t=6, t=9, and t=12 months (fig. 2).

In TRANSFoRm the QoL web questionnaire has an additional aspect. In WT 5.5 mobile applications are examined for the delivery of Patient Related Outcome (PRO). This work task will develop a mechanism for patients and clinicians to use the data collected from patients participating the GORD use case. It will record patient related outcome measures triggered by an index event in the EHR (consultation). These mobile applications can be mobile phones and smart phones used to fill out questionnaires used in the clinical trial. Patient Related Outcome Measures (PROM) will include Quality of life or patient evaluations of health care scores such as SF-36 Health Survey, SF-12 Health Survey and EQ-5D as well as symptom scores (categorized responses).

1. Patient access log in credentials and URL for web questionnaire
2. Patient access internet
3. Patient log in to URL
4. Patient fills out web questionnaire
5. Data is saved
6. Data is transferred to eHR (optional)
7. All steps, except obtaining log in credentials, are repeated for each follow-up

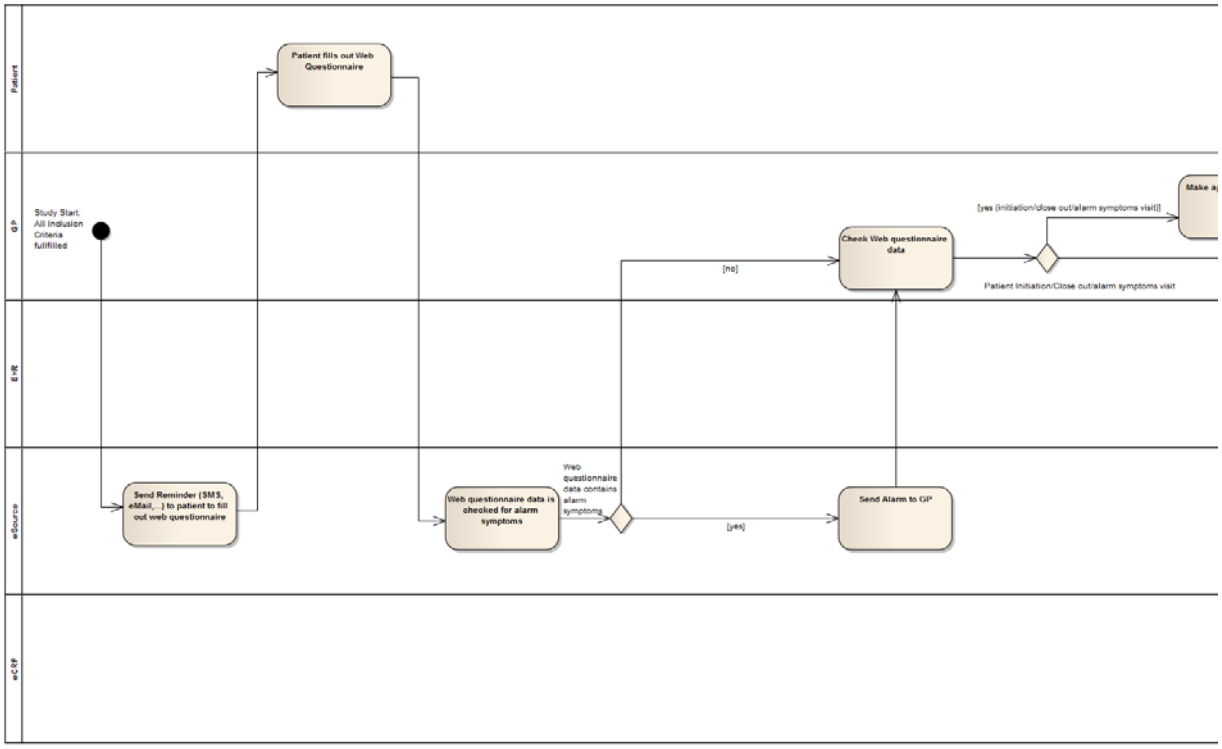
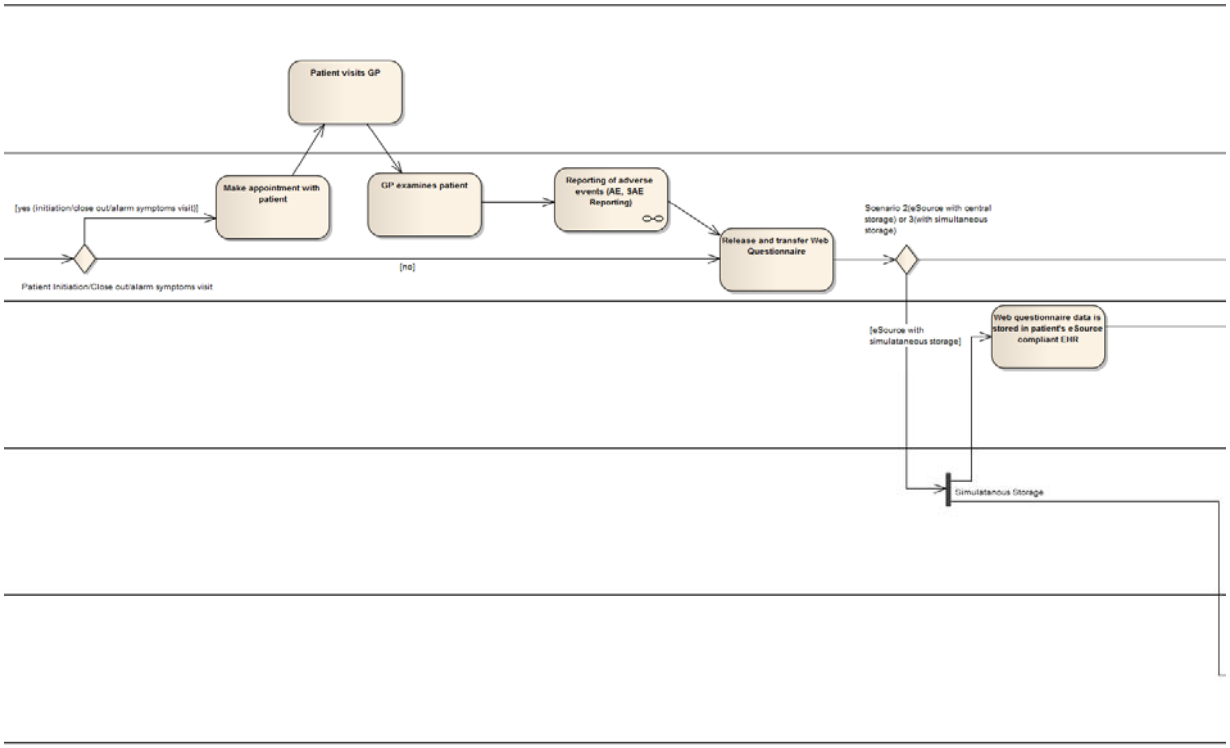
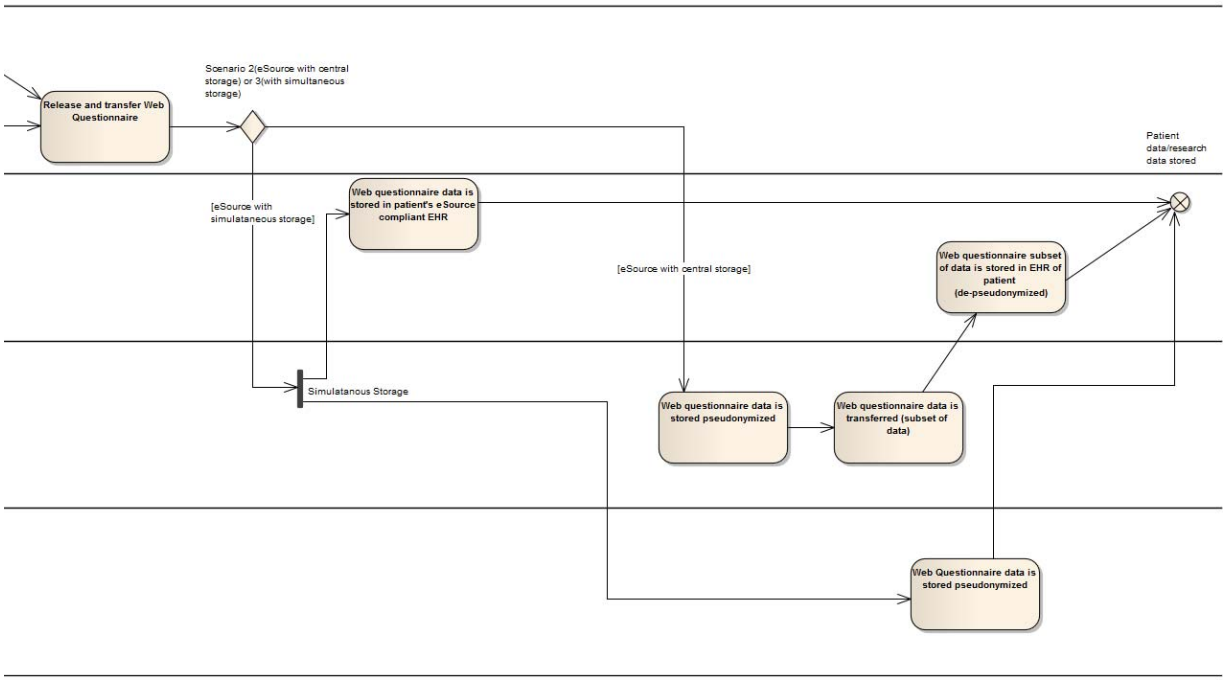


Fig. 7 Data collection with web questionnaire, triggered by alert (alarm)



(continued)



(continued)

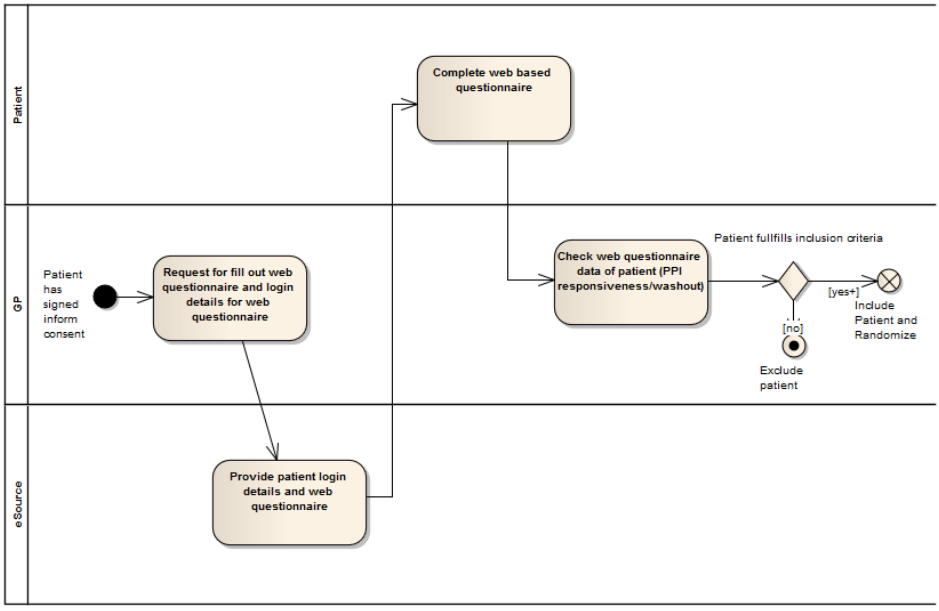
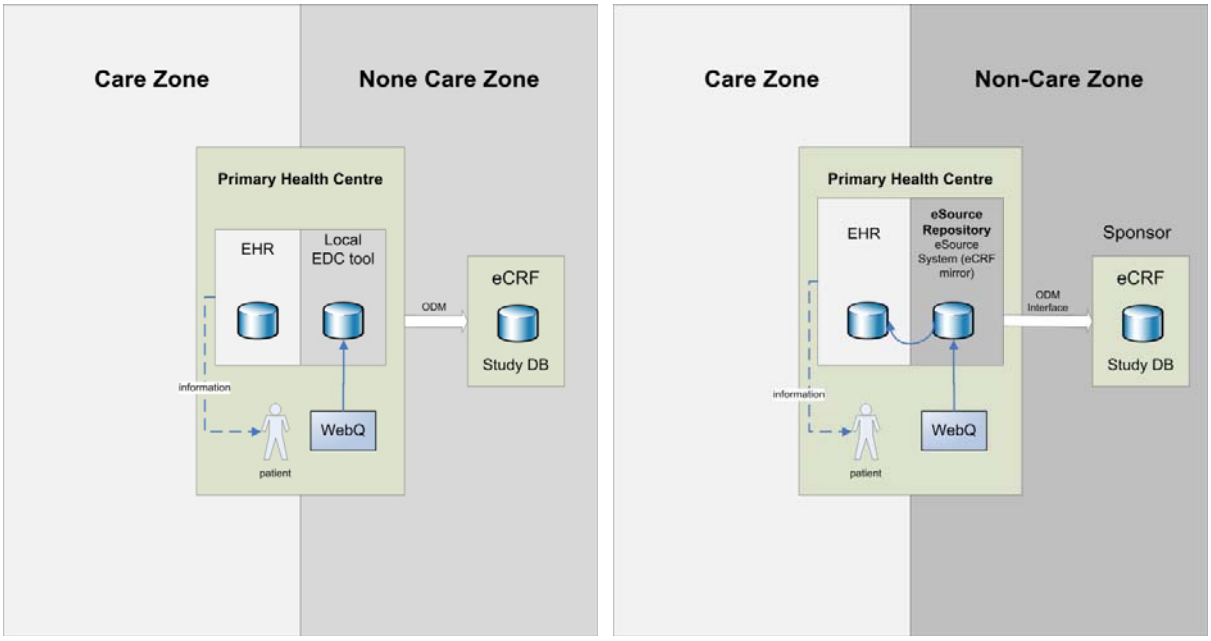


Fig. 8 Data collection with web questionnaire during first visit (t=0)

Because of the use of eHR, eCRF and web questionnaires for data collection, questions about the role of eSource data arises. Because quality of life data provided by patients using web questionnaires belong to the RCT, it is necessary to include this data into the eCRF and this makes it part of the study data base. This approach has several consequences.
According to the “Reflection paper for electronic source“ [12] by the GCP inspectors group, the sponsor shall not have exclusive control of a source document: “If source data are captured and entered directly into a web-based system without first being captured to paper or other local source and all data are stored in a central server that is not located at the investigator site, the sponsor should not have exclusive control of the source data. This requirement can be achieved by having a copy of the source data retained at the investigator site in addition to the record maintained on a central server. This is not always achievable. Procedures, system controls and contracts/agreements need to be in place to ensure that the sponsor does not have exclusive control. The sponsor of a study remains ultimately responsible for the quality of the study data and for ensuring that systems and processes are in place to protect this quality. Part of the process involved in achieving this may involve the use of service providers who furnish the hardware and may manage the software and data receipt and storage. “
This is also one of the requirements of CDISC eSDI [13]. According to CDISC eSDI, there exist four basic scenarios for using eSource data in clinical trials with eCRF. Three of these scenarios can be easily applied to the GORD Use Case (fig. 9). Scenario 2 and 3 were evaluated as most suitable for TRANSFoRm.



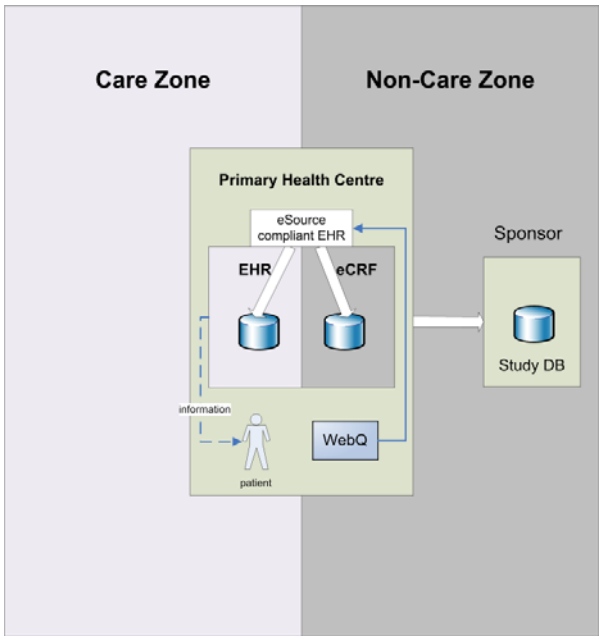


Fig. 9 eSource scenarios 1 (source at site), 2 (eSource repository) and 3 (eSource compliant EHR) and their location in the care and non-care zone of the privacy framework

Scenario 1 refers to the scenario „Source at site“ from the CDISC eSDI document [13] (p. 30). Here, source data (eSource) is maintained at investigative sites under the direct control of the investigator. In this scenario, the data from eSource technology (e.g. eDiary, eCRF or eData Collection Instruments) are sent directly to the principal investigator site. Storage of data is performed primarily locally; data may be subsequently transferred to the sponsor (study data base).

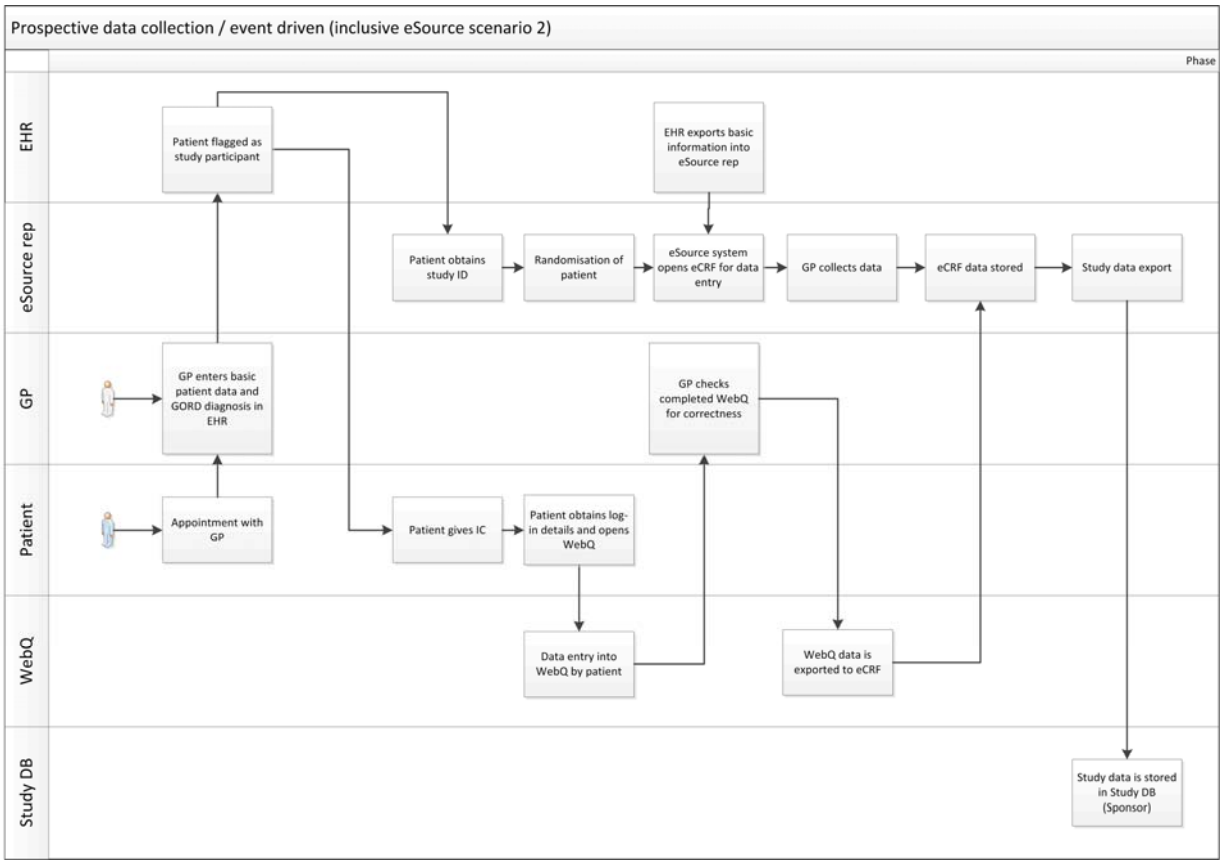


Fig. 10 Prospective data collection / event driven (inclusive eSource scenario 2)

Data collection workflow between scenarios 2 and 3 differ because of different functionalities of the eSource repository and the eSource compliant EHR. The information model must be conform to both scenarios. In our use case the patient enters data into the web questionnaire (WebQ) with an EDC tool, which is then transferred directly to the local database within the primary health centre (PHC). The GP is able to check the data and has the option to include the data (or some of it) in the EHR, provided this is a valid procedure. WebQ data is integrated with the eCRF (fig. 10 and 11).

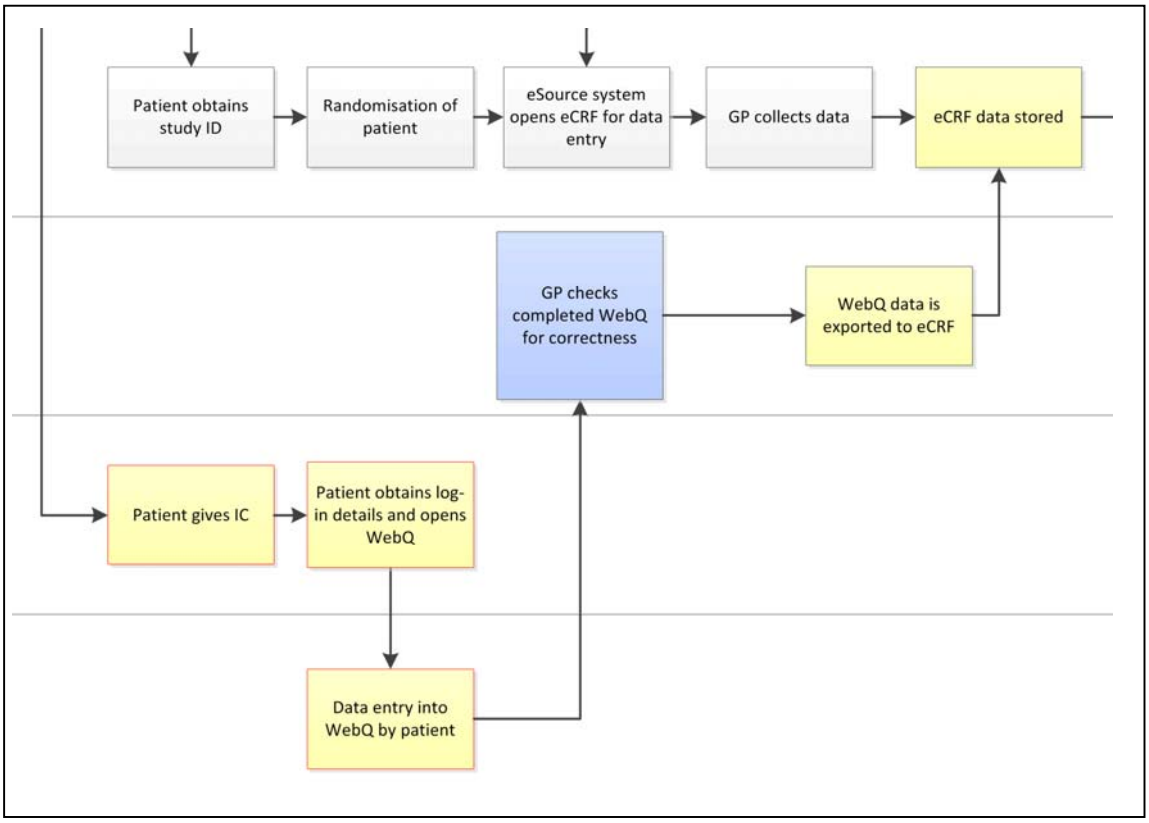


Fig. 11 Relationship of GP and patient during using the web questionnaire for data collection according to eSource scenario 2 (cut-out from figure 10). Yellow: steps concerned with the completion of the web questionnaire after recruitment for clinical trial, blue: data check step by GP.

Using an adequate interface, the data are transferred to the central eCRF and included in the study database. For this procedure an already available eCRF can be used. Nonetheless, eSource scenarios 2 or 3 may apply to this workflow.

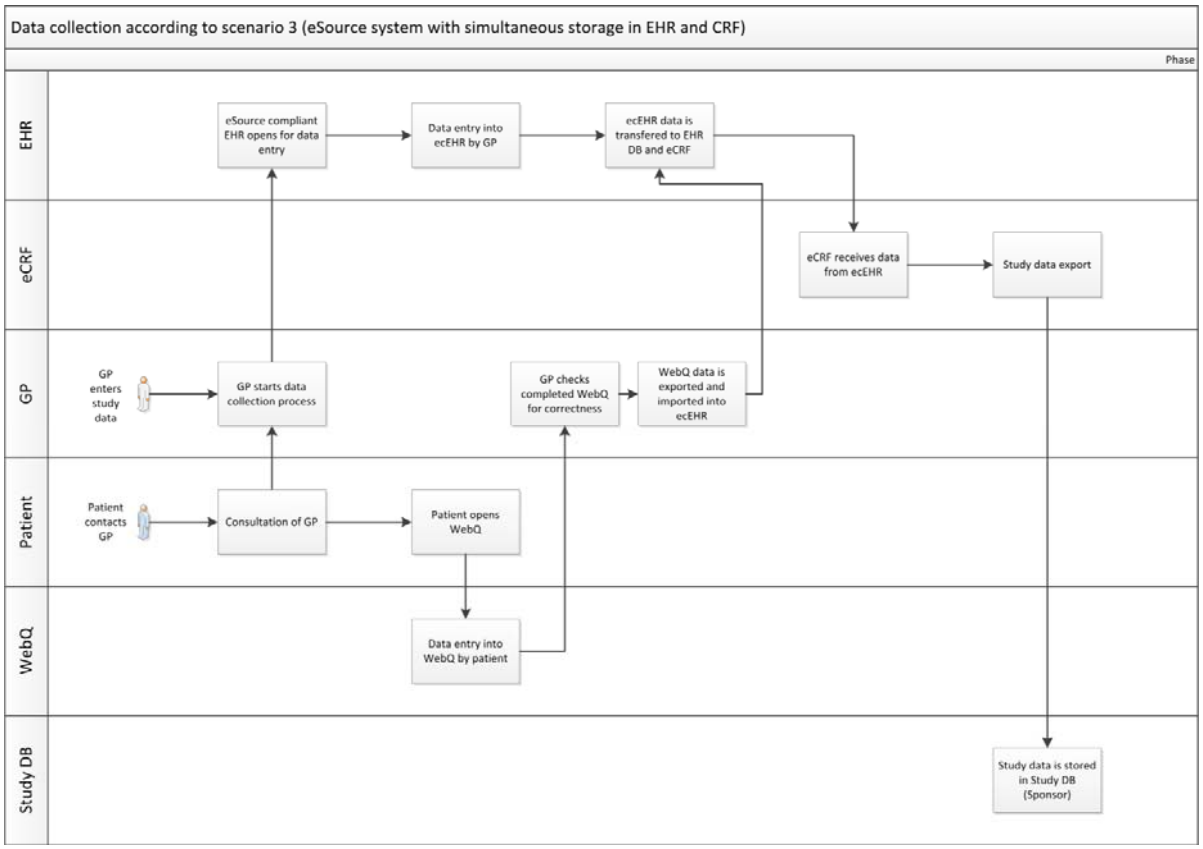


Fig. 12: Data collection according to scenario 3 (eSource system with simultaneous storage in EHR and CRF)

Scenario 3 (fig. 12) may not be the typical scenario for TRANSFoRm, but it may be applied in individual cases (e.g. in case a local data warehouse is used). The difficulty arises from the distributed storage of the web questionnaire data, making it necessary to transfer data from the different local databases to the central eCRF and to update the data. Problems may arise because of resistance of stakeholders concerning input of research data into the eHR (Source compliant eHR), the necessity to involve EHR vendors, the necessity to provide interfaces to different EHR products and the issue of pseudonymisation / anonymisation of data.

Use case 4: Randomization

Randomization of treatment after the patient has fulfilled all inclusion criteria and has given informed consent (fig. 13). The GP requests a randomization number.

1. GP reports having a patient fulfilling all inclusion and no exclusion criteria (includes informed consent) to system and asks for randomization
2. System generates a valid randomization
3. System reports randomization results to GP
4. GP prescribes drug according to randomization.

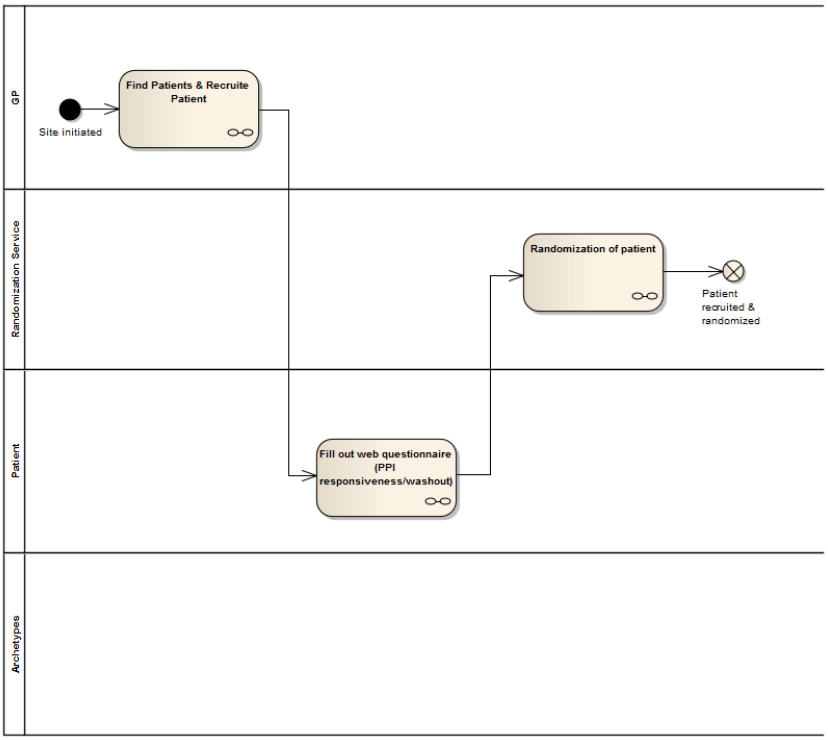


Fig. 13: Relationship between completion of web questionnaire and patient randomization

Use case 6: Extract information from databases

This general type of use case covers the extraction of information about selected patients with GORD (for finding patients with this characteristic see use case 1). The precondition is the availability of suitable data and the easy access for the researcher for the selection of variables to extract data from care databases. After the researcher has found patients to be included in the study, he selects what data need to be extracted (fig. 14). Data of defined variables are extracted for selected patients.

1. Researcher decides what kind of data are wanted for study
2. Researcher checks availability of data
3. Researcher selects what data to extract
4. Data is extracted to external database including personal identifiers for the linkage (next use case)

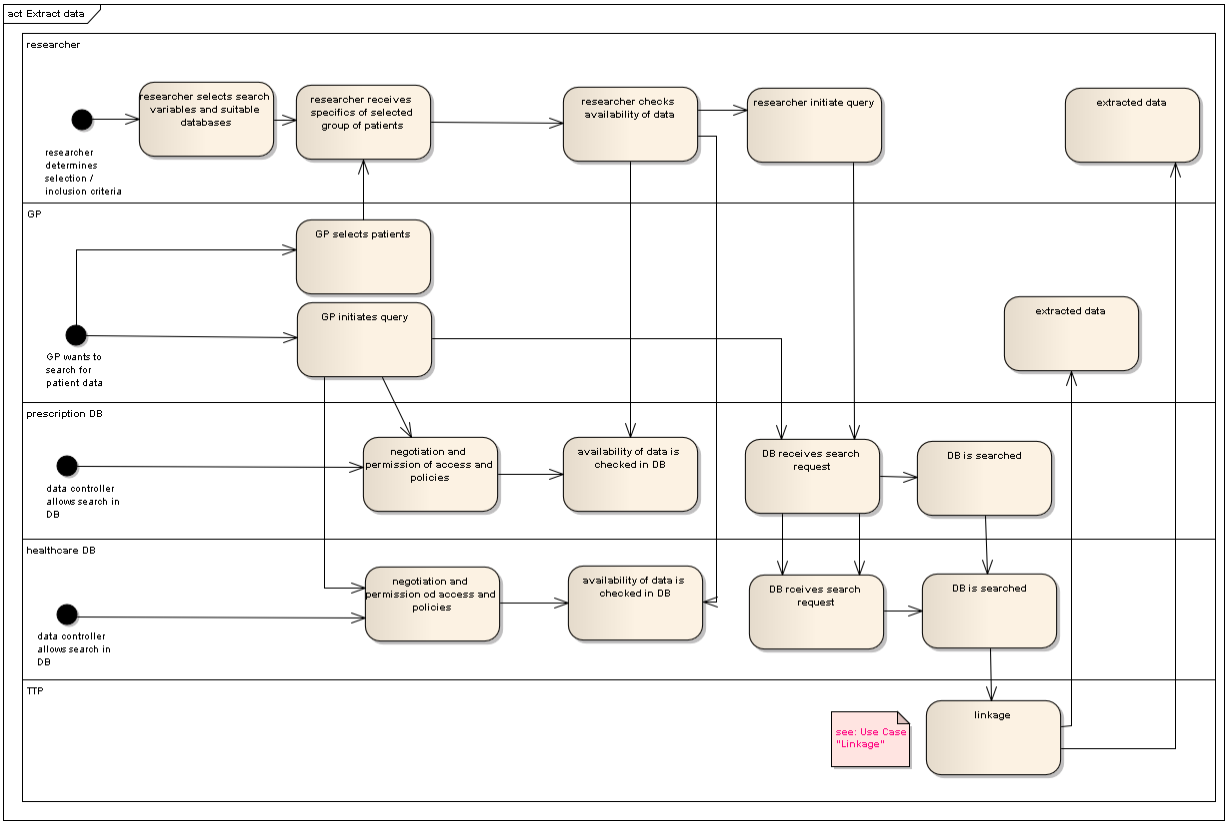


Fig. 14: Workflow for the extraction of data from prescription database and clinical care database.

Use case 7: Linkage of extracted data (general), including anonymisation

It may be the case that data from different care databases need to be linked using special identifiers for linkage. In the case personal identifiers are used, the linkage must be performed by a trusted third party (TTP). This can be done in different ways. The use case description document [1] demands that selected data is sent to the TTP where personal identifiers are removed. In general, TTPs act as administrator of pseudonyms; thus they do not receive the medical data. According to the data privacy framework of TRANSFoRm [14] the removal of identifiers is done by the data controller using a privacy filter before exporting data. Medical data are passed through to the receiver, whereas the corresponding pseudonyms are sent to the TTP.

1. Selected data has been extracted including unique personal identifiers
2. The selected data is sent to trusted third party
3. Trusted third party link the data
4. Trusted third party removes personal identifiers.
5. The new database is sent to researcher and stored.

The storage of the created linked datasets / databases, that contains no identifying information, is the last step. It may be for data protection reasons not possible that a linked database is sent to the researcher, but that the researcher obtains access to a linked database under the control

of a data controller.

Use case 8: Storage of data (general)

The data is in a format that easily allows for statistical analysis

1. Extracted data are transferred and stored available for researcher.

Combination of sub-use cases

All sub-use cases were combined and additional elements were added to generate a complete and feasible workflow description as basis for the information modelling (fig. 15 and fig. 16). The activity diagram covers data extraction from care data base and prescription data base as well as data collection according to eSource scenarios 2 and 3. For the combined activity diagram the processes depicted in the sub-use case diagrams were adapted and slightly changed and assembled to a single large diagram (fig.15).

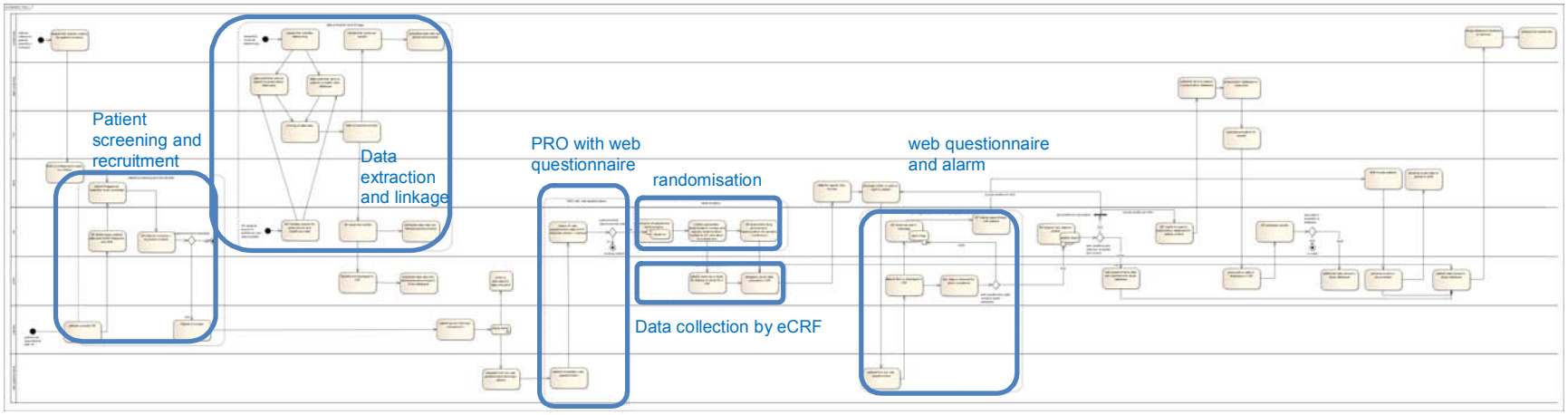


Fig. 15 Information and process workflow of the complete GORD use case (framed are the sub-use cases), PRO = Patient Reported Outcome

Activity diagram of the research and information workflow in the GORD use case

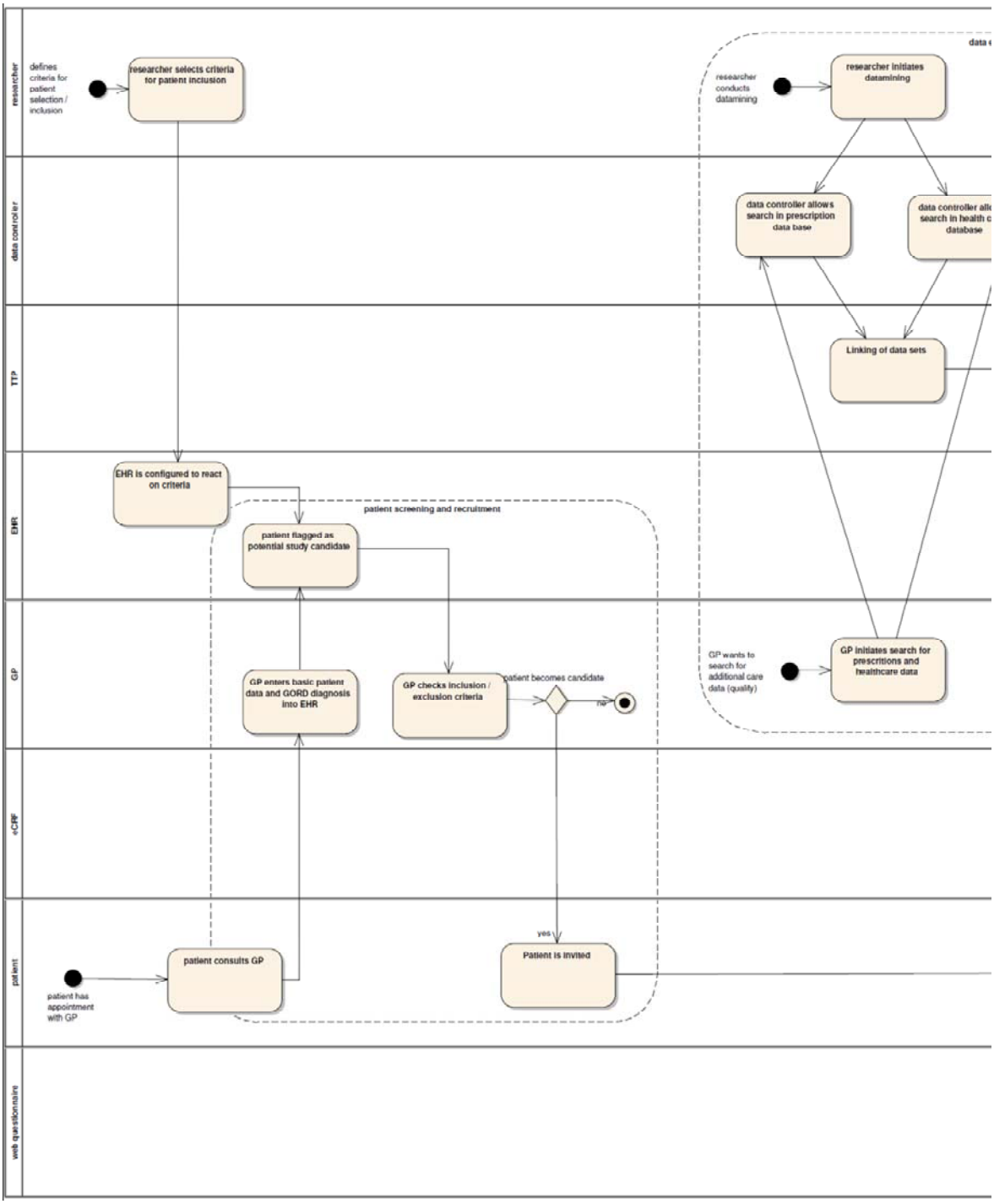
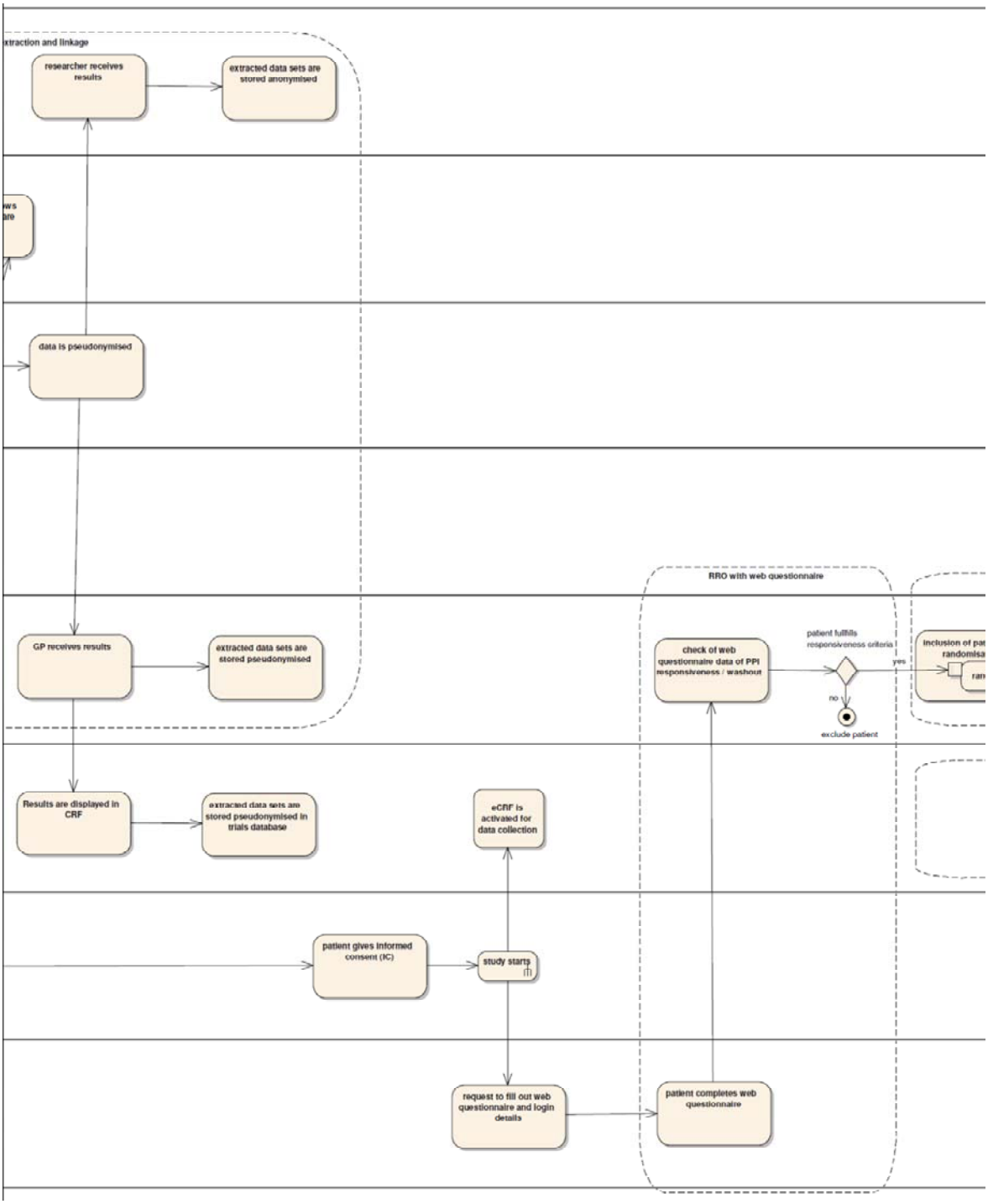
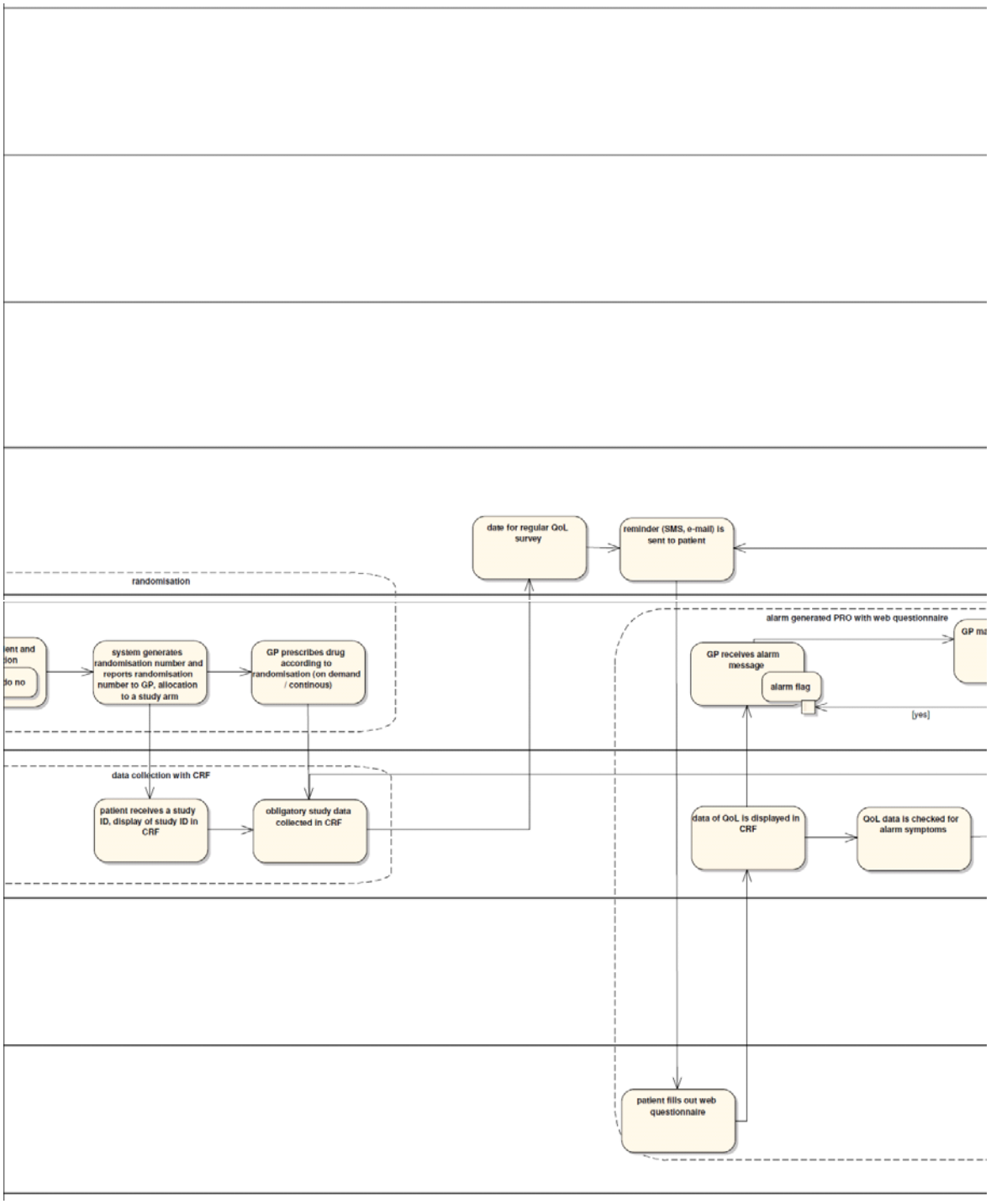


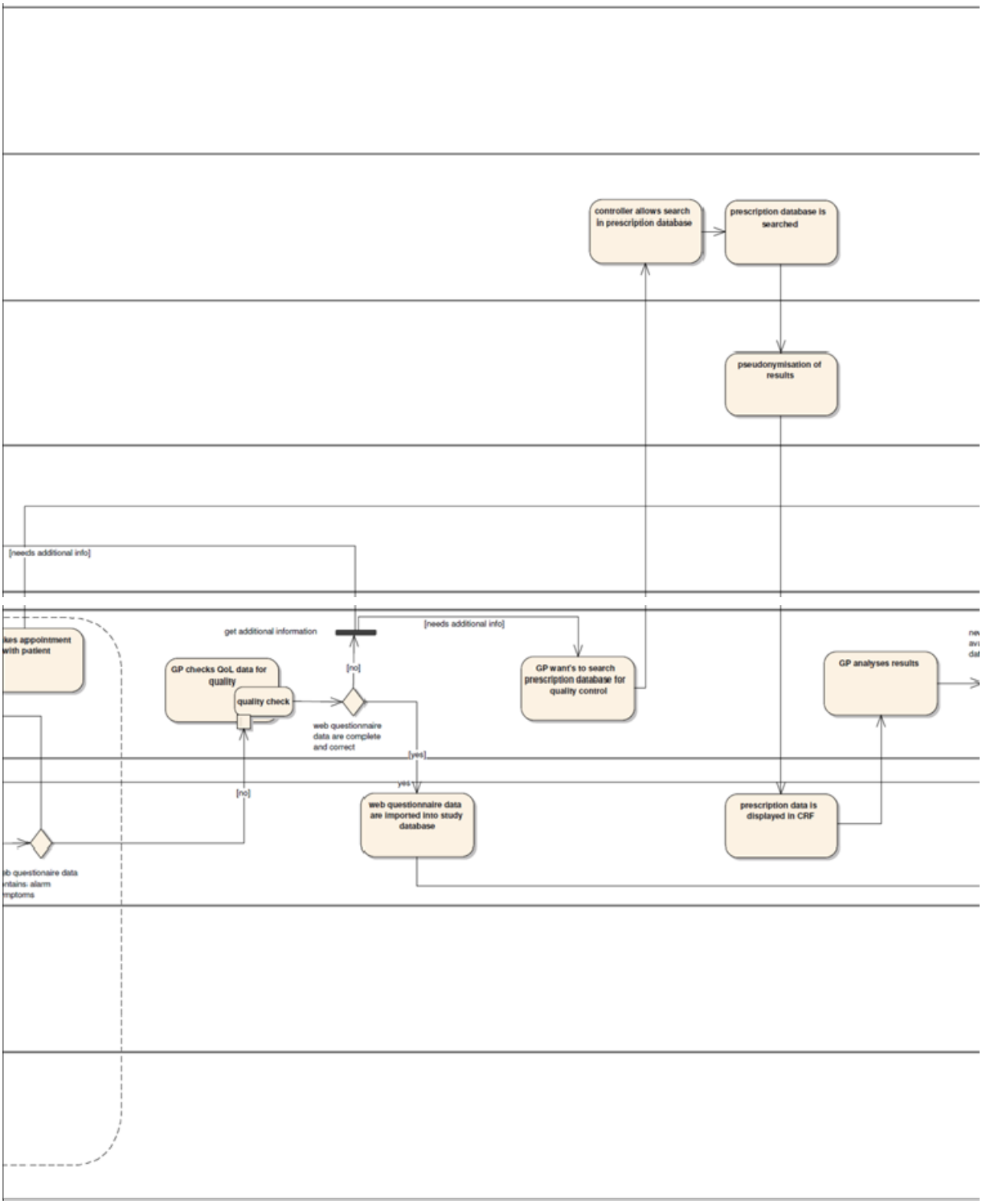
Fig. 16 Activity diagram of the research and information workflow in the GORD use case


researcher receives results
extracted data sets are stored anonymised
data is pseudonymised
GP receives results
extracted data sets are stored pseudonymised
Results are displayed in CRF
extracted data sets are stored pseudonymised in trials database
eCRF is activated for data collection
patient gives informed consent (IC)
study starts
request to fill out web questionnaire and login details
patient completes web questionnaire
RRO with web questionnaire
check of web questionnaire data of PPI responsiveness / washout
patient fulfills responsiveness criteria
yes
no
exclude patient

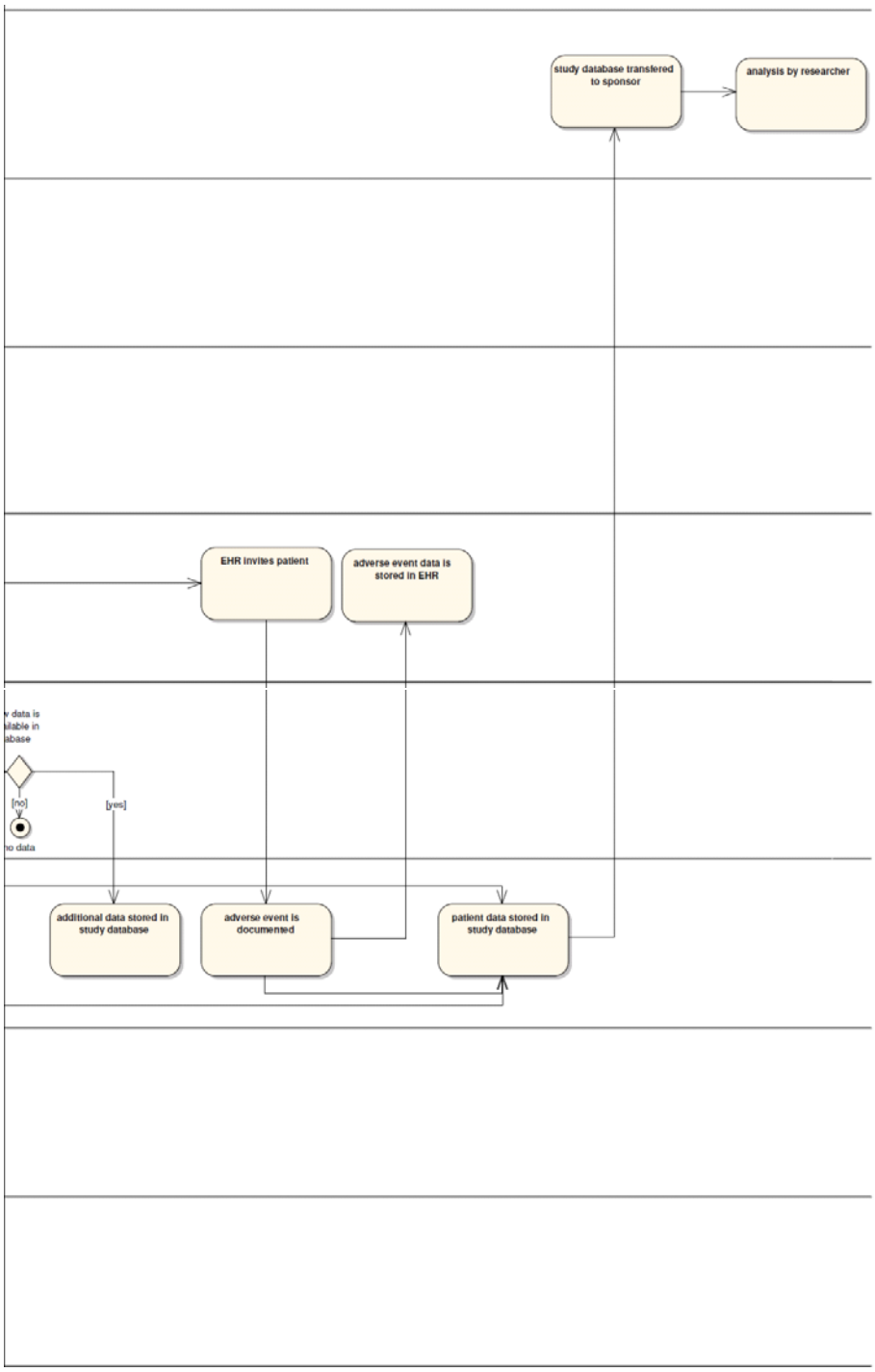
(continued)


date for regular QoL survey
reminder (SMS, e-mail) is sent to patient
randomisation
alarm generated PRO with web questionnaire
system generates randomisation number and reports randomisation number to GP, allocation to a study arm
GP prescribes drug according to randomisation (on demand / continous)
GP receives alarm message
alarm flag
[yes]
data collection with CRF
patient receives a study ID, display of study ID in CRF
obligatory study data collected in CRF
data of QoL is displayed in CRF
QoL data is checked for alarm symptoms
patient fills out web questionnaire

(continued)


controller allows search in prescription database
prescription database is searched
pseudonymisation of results
[needs additional info]
get additional information
[needs additional info]
GP checks QoL data for quality
quality check
[no]
web questionnaire data are complete and correct
GP want's to search prescription database for quality control
GP analyses results
[yes]
[no]
yes
web questionnaire data are imported into study database
prescription data is displayed in CRF

(continued)


study database transfered to sponsor
analysis by researcher
EHR invites patient
adverse event data is stored in EHR
[no]
[yes]
additional data stored in study database
adverse event is documented
patient data stored in study database

(continued)

6. Mapping of the GORD use case to PCROM /BRIDG

As described earlier, a three level modelling approach was used and adapted to the needs for primary care research. After developing activity diagrams for all GORD sub-use cases, a single workflow description of the entire GORD use case was developed by assembly of the sub-use cases. These activity diagrams were then used for a "walk through" relating all objects and the processes of the known information models (BRIDG, CTOM, CDISC) and especially to PCROM. The relevance of different assignments was discussed, and it was considered if existing entities and elements are appropriate, in terms of classes in the primary care domain.
Our use case modelling process identified both information required at each stage of the research process and details of the workflow. During this process, it became clear to us that the existing information models only insufficiently incorporate primary care research of the workflow requirements into their models. Through the modelling process we developed a complementary set of information objects which have a conceptual and relational connection with the workflow activities at the overlapping area between care and research (e.g. alerts (alarms), appointments, reminders, web questionnaire,…). The concepts of appointment, alert, etc. could also have importance for the information flow for the planned EHR based decision support.

Use Case activity diagram

The following GORD use cases were modelled and incorporated into the complete clinical trials reference model: (1) find patients for an RCT, (2) CRF (data compulsory for GP to record), (3) Quality of Life data (PRO), (4) randomizing patients, (5) extracting information from an EHR, (6) linkage of data (anonymisation), and (7) storage of data. The different roles of EHR, eCRF and web questionnaire were integrated with an eSource concept [13], analyzed for data protection risks and the eCRF was assigned the role of a central data hub. This was done not only as a requirement for the use case, but because this approach corresponds to the FDA way of thinking about the role of the eCRF in clinical research [11]. Only scenario 2 "eSource System Provider" and scenario 3 "eSource System with simultaneous storage in EHR" [13] turned out to be suitable. The resulting classes of the research model extended the PCROM information model of primary care research.

Discussion of modeling of the GORD use case

The first step is the search for patients and their recruitment. Here the care area and the EHR play a major role. Possible trial participants can be identified by alerting (EHR), diagnosis and search in eHRs. Data collection can be on defined time points and on-demand. On-demand data collection is triggered by alerting.
Data exchange between eHR, eCRF and web questionnaire is required. Because the technical

realization is not yet decided the information model should comply to both eSource scenarios 2 and 3. The importance of the care area covering the interaction between patient and GP and data collection with the EHR will make it necessary to include care specific elements in the information model.
Following distinct information objects were identified, analysed and mapped to PCROM as well as to BRIDG (table 2): AE, alert (alarm), appointment, close-out visit, consultation, CRF, data check, database controller (data controller), eHR, eligible patient, examination, GP (family doctor), GP visit, inclusion and exclusion criteria, informed consent, invitation of patient, medicinal product, patient, patient identifier (patient identifying information), patient initiation, patient recruitment, prescription, randomisation, reminder, research question, researcher, responsiveness, search, symptoms, TTP and web questionnaire. Several new concepts have to do with the interaction of GP (family doctor) and patient (e.g. appointment, examination, invitation of patient, patient initiation, prescription and several concepts are part of the TRANSFoRm data model and have little importance for the information model (e.g. data check, database controller, data controller, eHR). After mapping of the information objects to PCROM and BRIDG only 12 unique concepts were identified (table 3): reminder, alert / alarm, TTP, web questionnaire, GP (family doctor), invitation of patient, prescription, appointment (incl. GP visit), responsiveness (in EligibilityCriteria), patient screening, eCRF (Entry) as high-level unit, and consultation.

Relationship and relevance of GORD Use Case objects for PCROM and BRIDG

Analysis of the GORD use case activity diagrams generated a number of information objects that were analysed according to their relevance as class objects in PCROM and BRIDG. It was examined to determine if a new information object is already included in a class or in a relation, and what its relation to existing classes may be. The notion -> refers to the affiliation of an object to the data model, rather than to the information model.

No.	Class / Domain Objects Use Case GORD	Comment / Description	PCROM objects	BRIDG objects
1	AE	is result of assessment, assessment activity	AE, AEreport	DefinedAdverseEvent, AdverseEvent, (AdverseEventActionTakenRelationship, AdverseEventOutcomeAssessment, AdverseEventOutcomeResult, AdverseEventSeriousness)
2	alert (alarm)	is event (defined notification), process activity / study event (e.g. for AE event), alert is reaction to assessment		DefinedNotification,
3	appoint-ment	is related to care (e.g. GP visit)		
4	close-out visit	(also for follow-up visit) is a study event in a study, consultation	study event	StudyActivity (DefinedActivity)
5	consultation	is related to care		
6	CRF	-> data model, display as form (?) covering observation activity, assessment activity and intervention activity		Document (DocumentAuthor)
7	data check	-> data model		
8	database controller	-> data model		

	(data controller)			
9	EHR	-> data model		
10	eligible patient	is person, is result of eligibility search = process activity	potential participant (singular)	
11	examination		examination (is observational activity)	PlannedActivity, DefinedActivity (?)
12	GP	is person or role (family doctor, health care provider)	GP investigator	HealthCareProvider
13	GP visit	is related to care		
14	inclusion and exclusion criteria		EligibilityCriteria	DefinedEligibilityCriterion, DefinedInclusionCriterion, DefinedExclusionCriterion
15	informed consent	activity (?), potential participant consents (=relation)	consented as relationship	StudySubject, StudyProtocol (not directly dispayed)
16	invitation of patient	is related to care	notification	
17	medicinal product	is product, (medicinal intervention as a new class, covers "product")	intervention (generic), (medical intervention)	StudyAgent (Product, Drug)
18	patient	is a person	patient	Subject
19	patient identifier (patient identifying information)	-> data model		
20	patient initiation	potential participant, intervention	intervention activity	DefinedProcedure
21	patient recruitment	is process activity, see: eligible patient, participation is result	participant (as result)	StudySubject (as result)

22	prescription	process activity in medical care (prescription data base -> data model)	activity	Activity
23	randomi-sation	is a study procedure, process activity	allocation	RandomizationBookEntry
24	reminder	is event (scheduled notification), process activity /study event		DefinedNotification, PlannedNotification, NotificationReceiver
25	research question	is part of study protocol	study purpose	StudyObjective
26	researcher	is a person	investigator	QualifiedPerson, ResearchStaff, ResearchOrganization
27	responsive-ness	is assessment of PPI responsiveness = assessment results, assessment activity	assessment results, assessment activity	StudyActivity
28	search	(=eligibility search?)		
29	symptoms	result of an observational activity	observation activity result	DefinedObservation result
30	TTP	organisation		
31	web question-naire	is active process, e.g. PRO, observation different from the activity of data input, here it is observation activity		DefinedObservation

Table 2: Relationship and relevance of Use Case information objects for PCROM and BRIDG

After screening for their relevance, following objects were left over. These objects seem therefore suitable to be included in PCROM.

No.	Object
1	Reminder
2	alert / alarm
3	TTP
4	web questionnaire
5	GP
6	invitation of patient

7	prescription
8	appointment (incl. GP visit)
9	responsiveness (in EligibilityCriteria)
10	patient screening
11	eCRF (Entry) as high-level unit
12	consultation

Table 3: Suitable objects to be included in PCROM

The modelling resulted in following changes in the PCROM model:

- Altogether 12 unique concepts were identified that are important for an extended information model and that do not belong to a data model but into the information model.
- Introduction of a new high-level unit (eCRF/Entry) that includes the PCROM objects Intervention, AssessmentResult, ObservationActivity, ObservationResult and AssessmentAcctivity. This unit represents the central role of eCRF as an information object.
- The object WebQuestionnaire was added and linked to ObservationActivity (for PRO and QoL).
- PatientScreening (including Wash out) was added and linked to InterventionActivity.
- TTP was added as a new organization. It includes PatientIdentifier.
- A care area was introduced to combine all objects that are located at the overlap between care and research areas of importance especially for patient identification and screening (e.g. appointment, reminder). This area includes: Reminder, GP, Invitation of patient, prescription, appointment and consultation).
- EligibilityCriteria were extended by Responsiveness.

7. The Diabetes use case

Introduction

A considerable burden for health care systems is caused by the large occurrence of death and disabilities arising from Type 2 Diabetes (T2D). An important aspect in health care is the minimization of patient risks of developing complications, like retinopathy, renal disease and coronary heart diseases. Recently, evidences for convincing associations of genetic variants with T2D as well as their diabetic complications have been described. Research is now trying to integrate such genotypic data with data about risk factors to improve the accuracy of health risk predications. The T2D use case tries to determine if well selected single nucleotide polymorphisms (SNPs) are associated with the development of T2D complications in populations, especially if an association exists between SNPs and variations in drug response to oral anti-diabetics.

To answer this research question, selection of patients and data extraction from various databases has to be performed. The main steps involve (1) selection of T2D patients from genomic databases, (2) selection of T2D patients with a specific medication history from primary care databases (including databases derived from EHR data) and extraction of information (medication, SNP, HbA_{1c}, time since diagnosis, …) from these databases. To compare data from different databases it is necessary that the data are linked at the individual patient level. The study population of the use case consists of patients with type 2 diabetes (T2D), that are older than 18 years, live in the EU and whose genomic data has been collected in accessible databases and whose medical data has been collected in a primary care register. Such databases do exist in Europe. For example, the Scottish GO-DARTS database already contains data of diabetes patients linked to their genotypes for specific SNPs. An inventory of primary care databases in Europe was developed in WT 1.2.
The first step of the use case is that the researcher gets an idea of the size of the research relevant population in the available databases. A kind of preview will generate counts of the number of patients compliant to certain inclusion criteria. In a next step data sets of diabetes patients will be identified, extracted and linked. The research analysis will run on the combined data sets to identify patients with genetic risk factors. This scenario may not be possible for all databases. Some databases may only allow access to data, but no extraction and export of data sets. The information model has to cover both possibilities.

General storyboard

According to the use case [1], researchers want to use the TRANSFoRm system to create a new database which enables them to answer their research question(s). An overview provided to the researcher should present criteria for selecting eligible patients, the variables that can be extracted, information on comparability (how criteria, variables are supported by available databases), the number of patients available when choosing a certain design, specific criteria or when linking data from different databases, information on linking possibilities for data from different databases including information on necessity to obtain informed consent for linkage, information on costs. Based on this information the researcher will choose the databases to use (to be linked), criteria for the selection of target populations, and the variables to be extracted. After access and linking of data has been authorized, the researcher can start the analysis process. As result a report on the process of selection and extraction is generated by the system and presented to the researcher. Finally the researcher will end the search process.
The DT2 use case is divided in a number of sub-use cases for modeling. Central for the information modeling is to consider the processes that deal with the selection of cases and the extraction of EHR data from health care databases. In these areas clinical care processes overlap with clinical research requirements. DT2 use cases are:

1. Present system options (e.g. preview counts / characteristics of available research data)
2. Authorize data extraction and linkage (e.g. conditions for the need for informed consent)
3. Select patients with type 2 diabetes
4. Extract information (of selected patients)
5. Reintegrate data from different databases at the level of individual patients
6. Present data ready for further analysis

The use cases were examined for roles and actors and the work flows were analysed. As result of this analysis the role descriptions of several actors were changed (table 2). Especially for data extraction and linking, roles and concepts for the TRANSFoRm privacy framework were introduced. In table 1 similar actors / roles in the GORD use cases were listed, e.g. recipients, researchers, data source. Because their descriptions differ between GORD and DT2 use cases, for the modeling a common new description was adopted. In contrast to the joint roles, several actors are relevant only for DT2 use case: data source advisor, authority.

No	Role in use case	Description	New role in modeling
1	recipient (institution/server)	• receives the requested data to which researchers have access, • receives data from	recipient (institution, database like NIVEL,…), temporal storage place for linked data sets,

		databases · receives linking information from the linker, • linking of data based on unique identifiers, • integrates variables from different databases and/or EHRs, • analysis environment to the researchers	may provide access to linked data sets, applies a privacy filter to the linked data sets before sending them to researcher
2	authority	authorization, designating an acceptable recipient	data controller (is responsible for data)
3	executive authority	• authorization of requests from researcher to extract and link data • executive authority consults with the participating data source advisors, • presentation of conditions / procedures under which the request will be authorized (e.g. informed consent, permissions,…)	data controller (is responsible for data), authorization of access and linkage
4	informed consent	• for linkage	necessity of an informed consent is under the responsibility of data controller
5	specific specialist	• consults about security or privacy	-
6	researcher	• accesses the recipient's data, • defines the population and the variables which need to be extracted, • indicates whether data should be linked at the	researcher

		level of individuals, • makes formal request to an "authority' for data extraction and linkage • undertakes steps to obtain an informed consent for linkage and / or asks permission to use certain data	
7	system	• reports to the researcher, presents data	system is represented by a dynamic data discovery service (Quality tool) and by the Query tool
8	data sources	• interact with the linker(s) in order to obtain a unique patient number • interact with the system in order to select patients and to extract variables, • release extracted information of selected to the recipient	data sources reside in different non-care sub zones, may offer services, like patient selection or data extraction
9	data source advisors	• protects the interests of the data sources, • provides information when specific questions are asked (concerning content, quality, privacy)	data controller
10	linkers	• manage identifiers and anonymization, • communicate with the databases,· communicate with the recipient	TTP and coreTTP

Table 4: Actors and roles in the Diabetes use case.

Governance

In DT2 use case governance issues are concerned with processes of data encryption, data transfer, data storage, data linkage, accessibility (who has access). In addition, the kind and number of linkers, where and how linking will be performed and how the accuracy of linking will be guaranteed may vary according to country. The roles of GP, researcher and data controller as well as the role of databases and TTP are described in the TRANSFoRm privacy framework [14]. In the framework the area of care and the area of care databases are describes as “care zone” and “non-care zone”. This zone model was applied to the modeling of the use cases.
Country/database specific linking could be done by a type of linkers where the TTP is holding the identifiers for a database. How many times and in which countries a linking process can be performed, how many TTPs and super-TTPs are involved has to be explored. To integrate databases from different European countries and to link data at the level of individual persons, legal and ethical regulations within the countries and within Europe as a whole should be followed. Necessary applications, assessments and approvals need to be obtained.
Attention points listed in the use case are: (1) authorization is needed for each researcher’s data request; (2) database and country specific regulations and policies have to be regarded (e.g requirements for an informed consent for data linkage); (3) for data transmission an agreement should be made between the data source and the recipient; (4) linkers can be seen as Trusted Third Parties (TTP) and the number of linkers, where they reside and how they will obtain information depends on the databases involved. (5) a track of the provenance of the data transmitted.
Guaranteeing provenance tracking is an important aspect for the use of care data for research purposes. In TRANSFoRm a service for managing provenance of eHR data obtained for research (WT 5.4) is essential for validating data quality and auditing compliance with the privacy framework TRANSFoRm. The Provenance service will capture provenance through logging of instructions by the data source clients and executives, including items such as timing of operations, authorisation of instructions etc. The framework will be responsible for capturing the system level provenance during the interactions with the middleware, while observing the minimal legal requirements for storing and retrieving provenance data. It will not enforce policies, and will only provide an audit trail with the capability to reproduce an analysis on demand. An important aspect is, that the information needed for the contextual provenance is normally not shared with the participants that are outside the data source organization, It is planned, that at a basic level, the following information should be captured: user information, time stamp of the request, authorisation certificate, preview query, preview response (counts/characteristics metadata), full query (selection), reference to the returned result, integration operation, reference to the integrated data set.

Use case 1: Present system options

In this sub use case the system represented by the Dynamic Data Discovery Service (and the Quality tool) is used to obtain basic metadata about the databases. For this the system needs a list of databases to search. There are two scenarios possible: (1) search is done on pre-specified data and (2) search is done newly each time. In the first case a register with all relevant metadata of all databases have to be created and maintained. Large care databases, like GPRD and NIVEL, make already considerable information on their websites available.
Presenting systems options means that after this process the researcher knows what data and with what quality is available and whether a linking is possible. The researcher has to know how many patients of the genetic database can be linked to the care data base (with clinical information). In addition to the information about the suitable databases and available variables, the researcher needs to know by whom and the linking will be performed. Here information should be provided by data controller and TTP representatives. The Quality tool may not provide this information automatically with the query.

1. Large care databases are not passive data stores but act as research service providers participating actively in research projects. NIVEL as well as GPRD offer a variety of services, including search and result presentation processes. Researcher wants to have an overview of the variables of interest (selection and extraction data elements) as available in the primary care and genomic databases and wants to learn about conditions and costs
2. The researcher initiates the system options procedure
3. The system presents the researcher with a list of the databases, included variables, etc.
4. Characteristics of these variables including information on comparability and quality information
5. The researcher explores system options
6. The use case ends

The first step is the search for suitable database contents (fig. 17). Using the dynamic data discovery service with the Quality tool, the researcher selects the databases and defines the search questions regarding the database metadata, e.g. availability of data, completeness of data, availability of policies, costs. These metadata are already available in a special repository, or have to be searched for in the databases. In addition it might be only possible that database provides make this data available on request (fig.18). All three possibilities are included in the activity diagram. Because these three scenarios have different implications, e.g. local interface / adaption / policies, they are presented in the workflow diagram in a general way without resort to technical specifics. The results of the search / request are displayed, enabling the researcher to select suitable databases for the patient selection query. In particular, provided results will enable him to judge about the possibilities of record linkage.

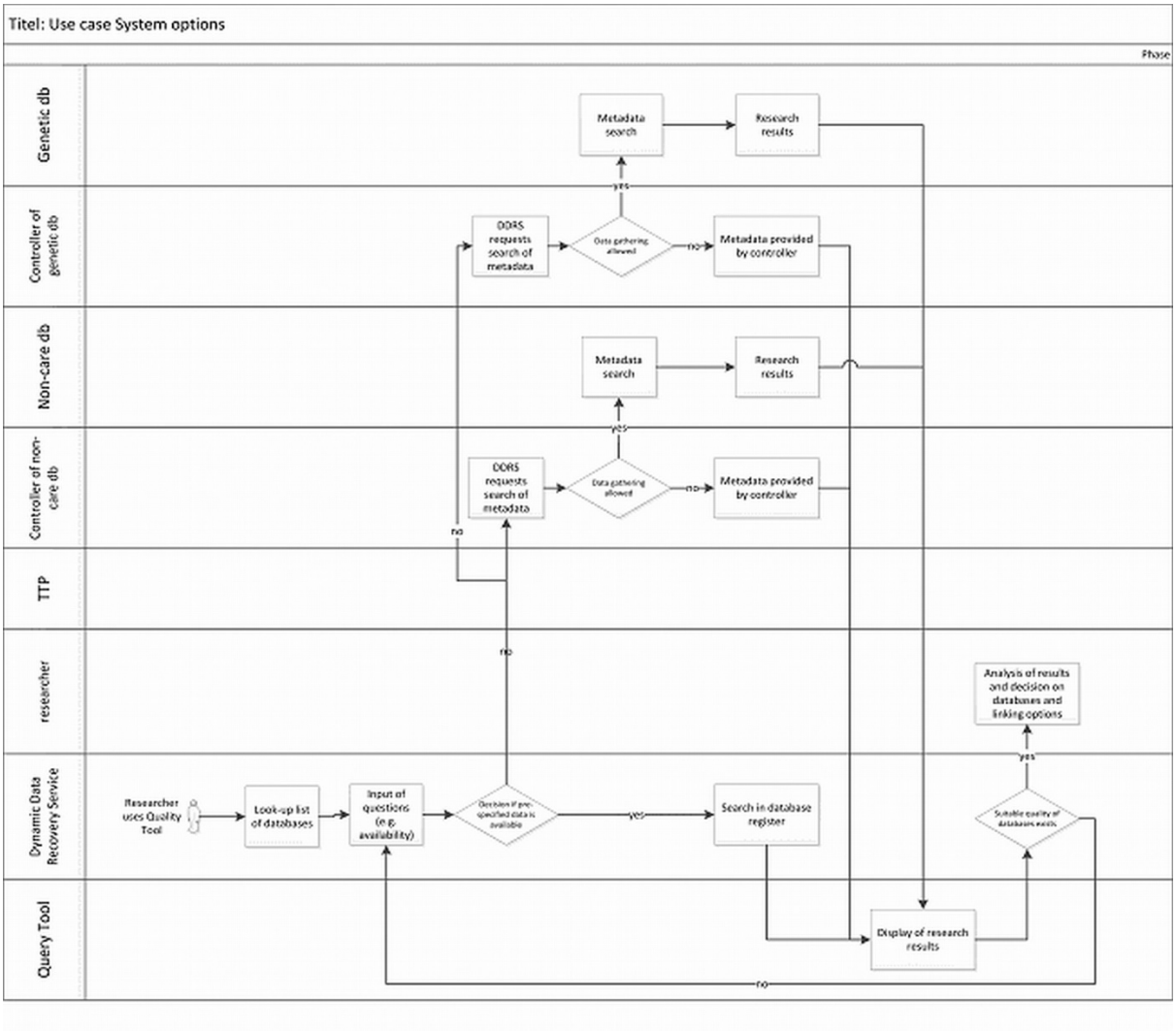


Fig. 17 Sub use case 1: presentation of system options (see complete activity diagram at the end of the report) (DDRS = Dynamic Data Discovery Service)

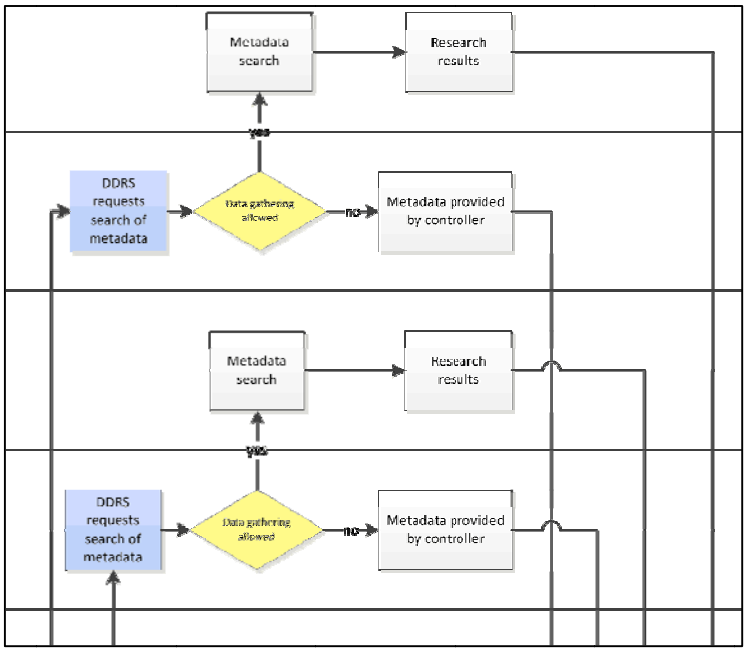


Fig. 18 Presentation of system options (cut-out of fig. 17): decision points (yellow) concerning the DDRS (blue) that decide if it is possible to obtain information about databases automatically (data gathering allowed = metadata search) or by metadata provided by the controller (DDRS = Dynamic Data Discovery Service)

Use case 2: Authorize data selection, extraction and linkage'

The Diabetes use case takes place entirely in the non-care zone of the zone model of the TRANSFoRm privacy framework [14] according to the definition of the use case. In agreement with this framework the authority is represented by the "data controller", a defined role of the EU Data Privacy Directive. The data controller is responsible for the data and decides who has access and what can be done with the data ("data processing"). All non-care databases have a data controller.

1. Researcher requests permission for patient selection, data extraction and linkage
2. Authority checks which conditions need to be fulfilled and presents these to the researcher
3. Requirements are fulfilled and presented to the authority
4. The authority responds to the researcher
5. The authority authorizes the start of the workflow for selection, extraction and linkage
6. The use case ends

In a second step the researcher starts the querying process using the Query tool. As a precondition the access to data, data extraction and data linkage must be authorized. In addition, it might be possible that only access, but no extraction or extraction without linkage may be permitted. Large care databases employ different policies in this regard: the Health

Informatics Centre (HIC) at the University of Dundee supports research through the collection and management of high quality data in a secure environment with strong data governance. For NHS projects a secure transmission of data from the NHS network to a data recipient is possible [15]. On the other hand, NIVEL prefers to give access to datasets and does not allow transfer of data to a researcher. Both possibilities are included in the workflow. After access / extraction data sets are approved by the data controller of the non-care database and the genomic database the T2D patient selection can start on the combined databases. Selection criteria are defined and the corresponding data sets are identified in the corresponding databases.

1. In case data extraction is allowed, marked patients sets are extracted and stored temporarily at the "recipient". The recipient can be an independent institution or one of the involved databases. It must be considered that TTPs don't store medical data, but only hold pseudonyms to link cases.
2. In case extraction is allowed, but linked data have to be kept by the recipient, the recipient can grant access to the linked datasets for analysis
3. In case neither extraction nor linkage is permitted, the databases might allow remote analysis of the selected data sets. Remote analysis allows a distributed analysis on the concerned databases and subsequent result merging and analysis (see: dataSHIELD [16])

Data aggregation through anonymous summary statistics from harmonized individual-level databases (dataSHIELD) provides a simple approach to examine pooled data. This is achieved via parallelized analysis and modern distributed computing and, in one key setting, takes advantage of the properties of the updating algorithm for generalized linear models (GLMs) [16]. This parallelized analysis is so performed that the only information passing back and forth between computers consists of short blocks of computer code defining the next analysis required, and low-dimensional summary statistics used in estimating the mathematical parameters of the model. Thus, these items disclose neither the identity, nor the characteristics, of individual study participants.

The non-care database as well as the genomic database has a data controller and a TTP. The corresponding TTP holds in each case the pseudonyms of the concerned patient data sets. The pseudonym might be something like a social security number (like in Sweden), or a random number. To link both data sets the pseudonyms of both databases must be mapped in a secure environment. This will be performed by an additional coreTTP (Custodix).
According to the TRANSFoRm privacy framework [14], genomic data are potentially identifiable. Therefore, an additional protection mechanism must be employed for the linked data sets that now contain genetic information. Therefore a privacy filter (e.g. an aggregation step) is used before the anonymised data are sent to the Query tool and the researcher. As an alternative, these anonymised data are not transferred, but are made available for remote analysis.
Once the linked data set is received, it has to be processed to be in a special format for scientific analysis by a tool of choice. In this way, patients with genetic risk factors are identified.

Use case 3: Selection of patients

In this sub use case the Query tool is used for the selection and flagging of patients in different databases. Two possibilities exists: (1) the query tool can access the non-care database and search for suitable patients; (2) the search is actually done by a tool specifically used by the corresponding database and the query tool act to transfer the search criteria and receives / displays search results. In a first step the number of eligible patients (counts) is displayed, in a second step the eligible patients are selected and marked.

1. Researcher wants to select T2D patients from different primary care and genomic databases
2. The researcher logs on to the system and initiates the selection procedure
3. The system presents the researcher with a list of selection criteria including information on availability in the different databases and quality information
4. The researcher chooses databases and selection criteria
5. The researcher 'initiates' a "preview" module to get an idea on the number of included patients with the selected criteria, databases (and cases and controls) - OPTIONAL
6. The researcher confirms or adapts selection criteria/databases
7. The system runs the "T2D selection module" in the databases of choice and selects eligible patients (flagged patients)
8. The use case ends

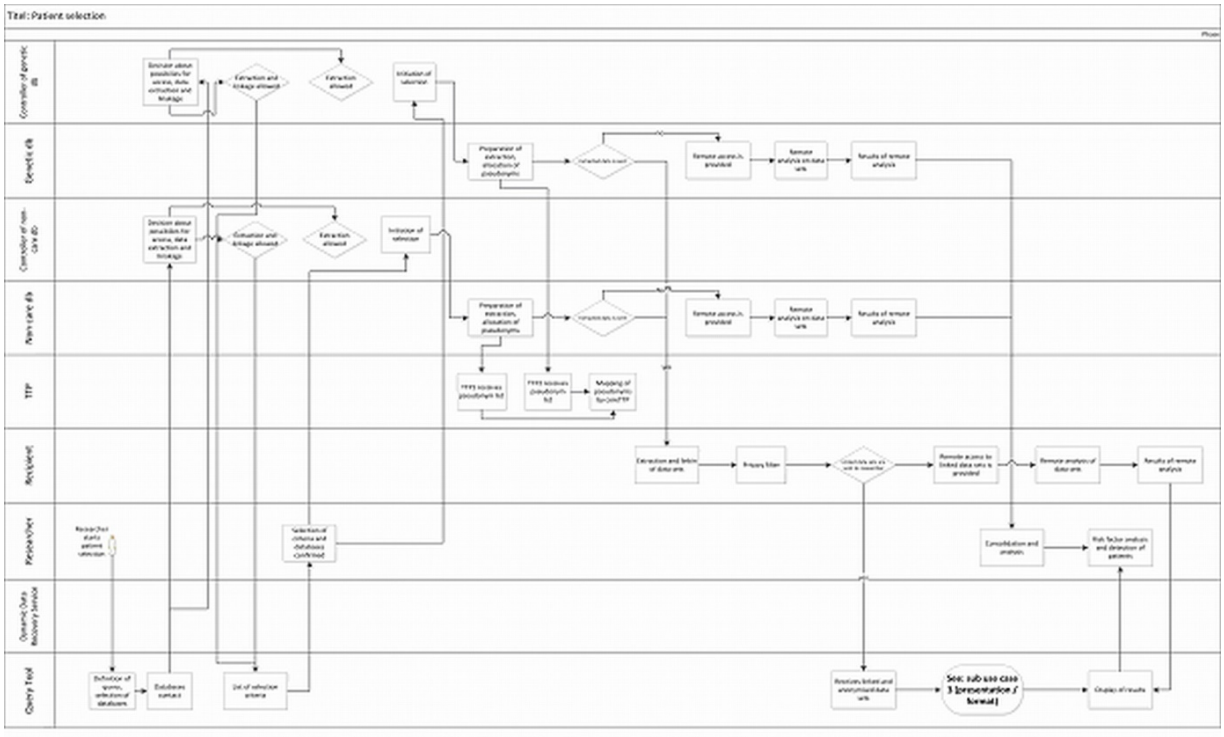


Fig. 19 Sub use case 3, 4 and 5 that are concerned with patient selection were combined (see complete activity diagram fig. 20)

Use case 4: Extract information

This use case description overlaps with use case 'select patients'. It must be considered that it might be not possible to extract data sets, but that the database will give access to data (see use case 3). In general it will be necessary to link information on the same patient from different databases (see use case 5).

1. Researcher wants to extract information for selected T2D patients from different primary care databases and genomic databases.
2. The researcher logs on to the system and initiates the extraction procedure
3. The system presents the researcher with a list of variables including information on availability in the different databases and comparability and quality information
4. The researcher chooses variables and databases
5. The researcher initiates the selection procedure by confirming the list of variables that needs to be extracted (and how they are defined based on archetypes)
6. The researcher 'initiates' "preview" module to get an idea on the number of patients for whom the selected variables are available (and numbers of cases and controls) - OPTIONAL
7. The researcher confirms or adapts (back to step 4) choice of variables and databases
8. The system runs the "T2D extraction module" in the databases of choice or eHR and extracts information requested.
9. The use case ends.

Use case 5: Link and reintegrate data

In this use case the „recipient" as a new role is introduced. Because the TTP will only store pseudonym lists, or mapping information, a storage place for the linked data sets is necessary.

1. Data sources transfer extracted information to the recipient. In contrast to medical information, primary keys for each patient may be transferred to and stored at a TTP.
2. The linker(s) provide(s) the recipient with the link between primary keys in the different data sources
3. The recipient links information at the level of individuals based on the information provided by the linker(s)
4. The recipient reintegrates variables from different sources
5. The use case ends

Use case 6: Present data

This use case adds an addition step to present the received data in a analysis-friendly format.

1. The recipient receives instructions on the data output format (structure of the new

dataset, output)

2. The recipient prepares the data
3. The researcher is informed the data is ready
4. The researcher can access the data
5. End of the use case

Combination of sub-use cases

After developing UML activity diagrams for all Diabetes sub-use cases, a single workflow description of the entire Diabetes use case was developed (fig. 20 and 21). The different roles of databases, data controllers and researcher were integrated with the concepts of the TRANSFoRm privacy framework (e.g. data controller, TTP, privacy filter). The initiator of the workflow is the researcher who defines the research question and executes the query. A special role of the GP in patient selection and recruitment as it was described in the GORD use case was not considered for the DT2 use case because the use case deals with an observational study and direct involvement of the care area is in this case not necessary. All sub-use cases were combined and additional elements were added to generate a complete and feasible workflow description as basis for the information modeling. For the combined activity diagram the sub-use case diagrams were slightly adapted.

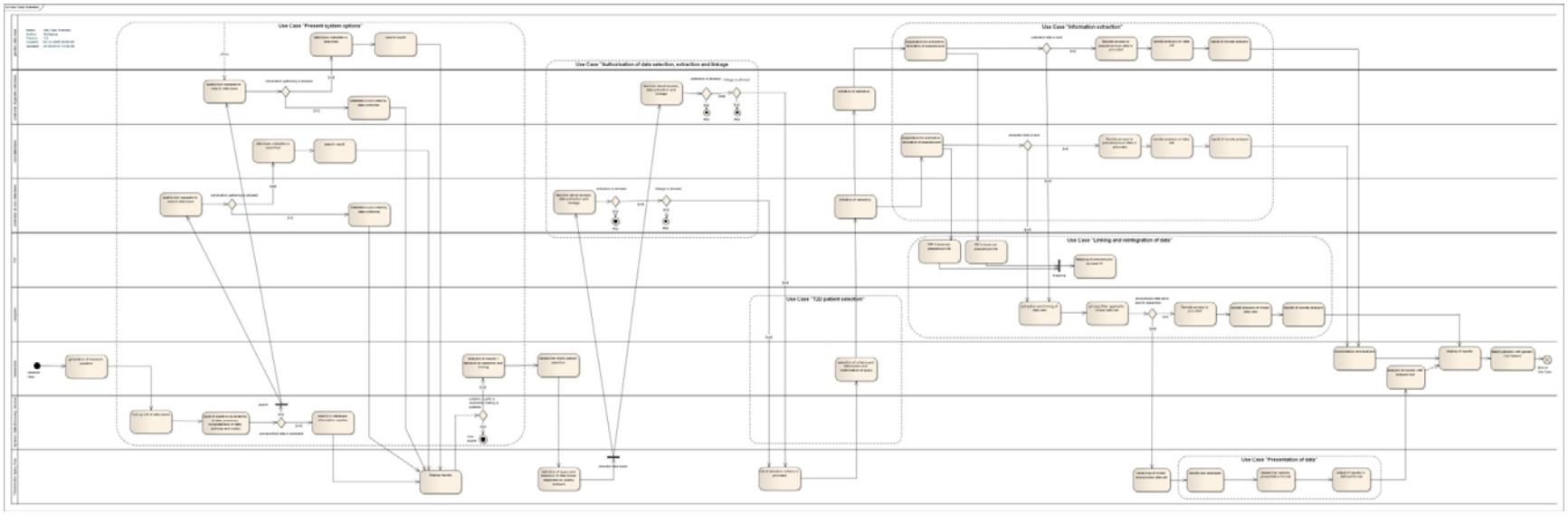


Fig. 20 Complete workflow of Diabetes use case depicted as activity diagram

Activity diagram of the research and information workflow in Diabetes use case

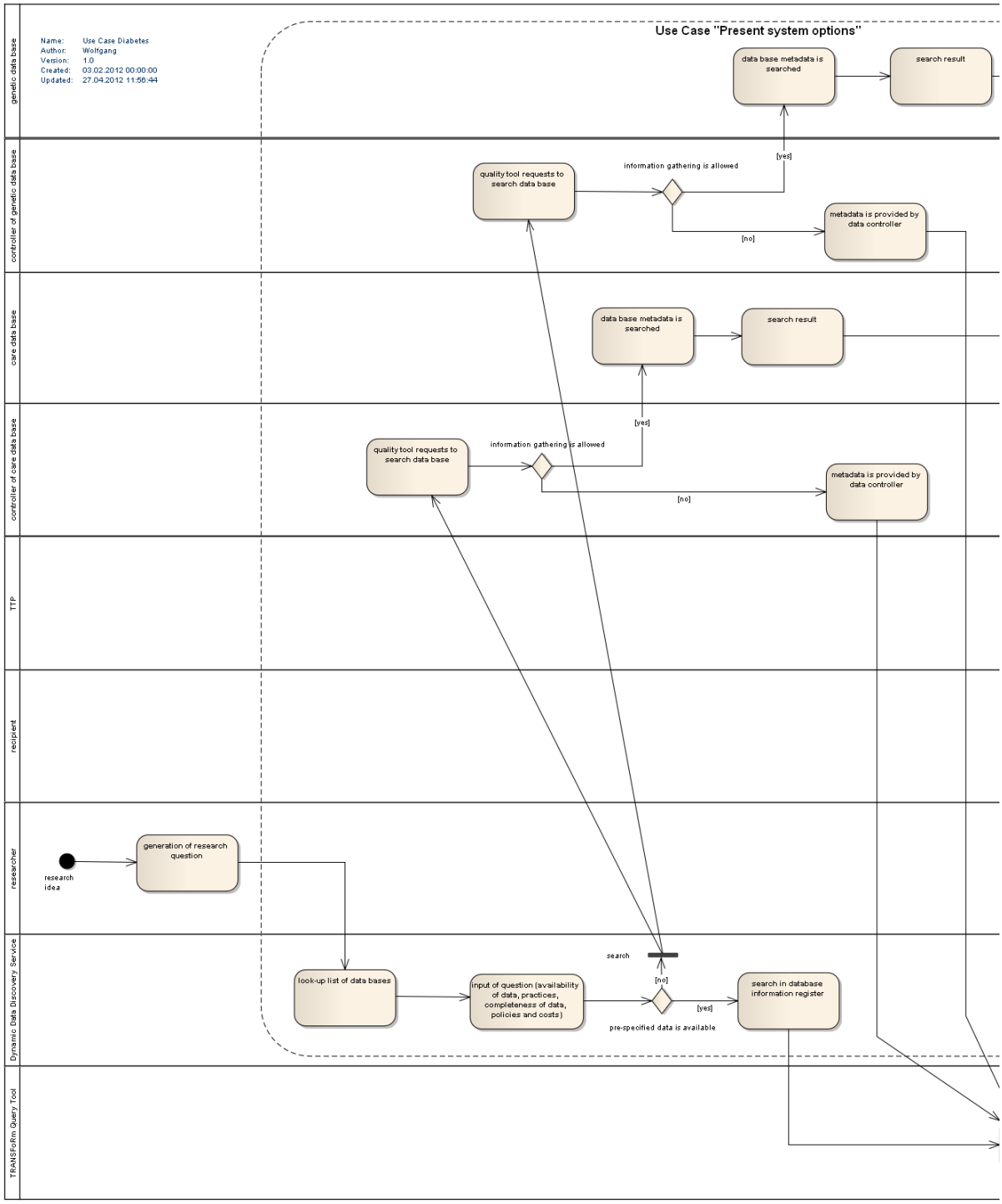
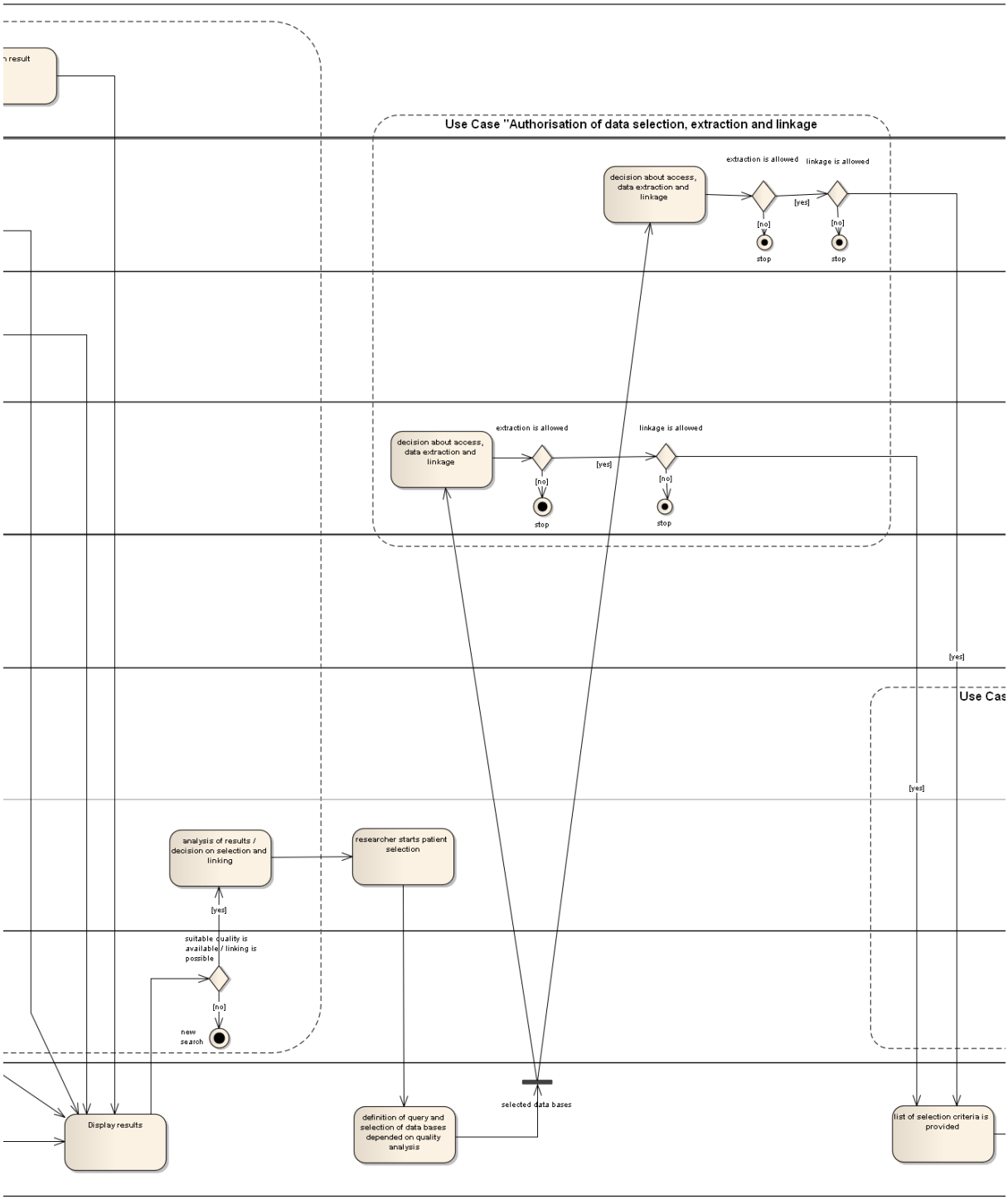
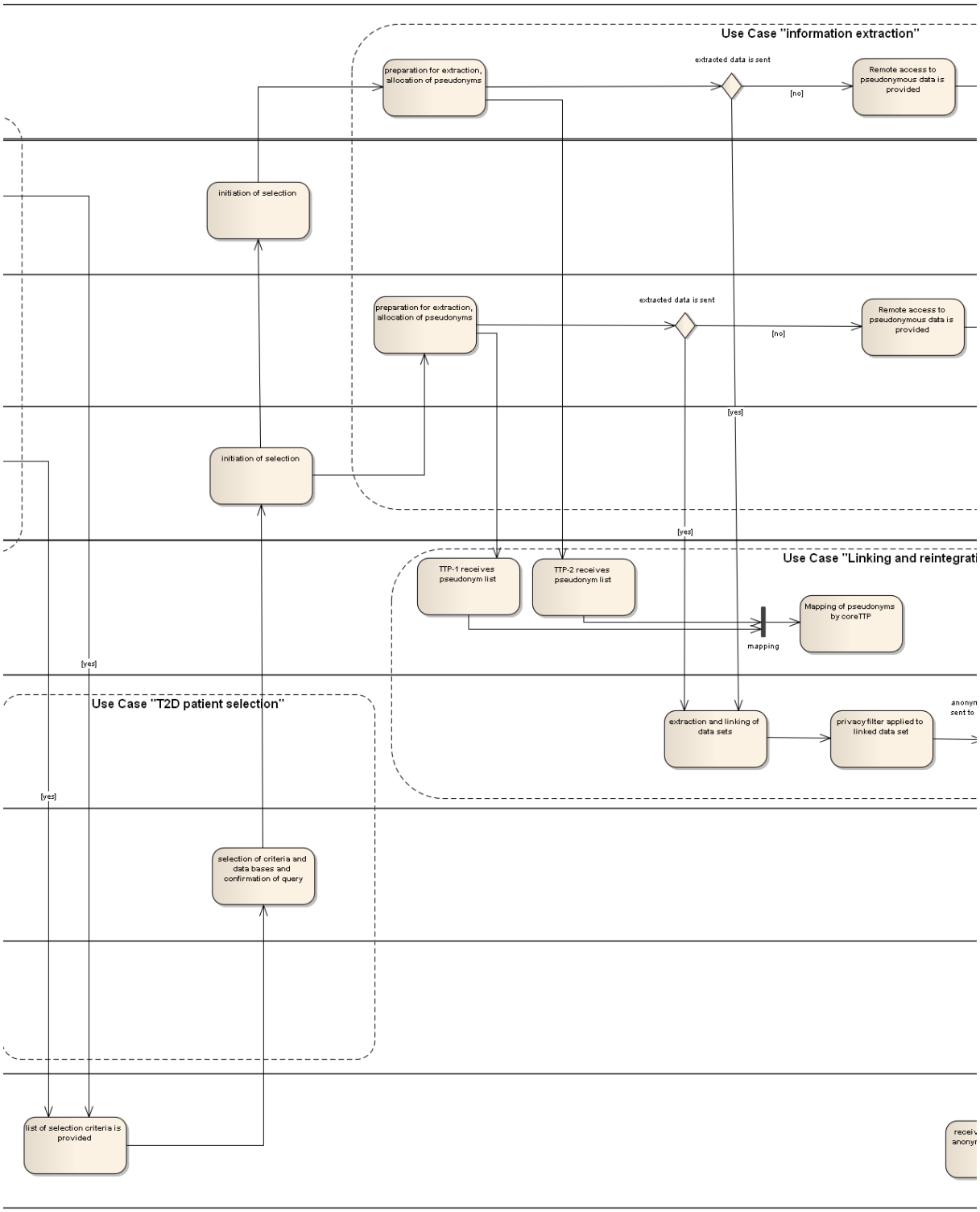


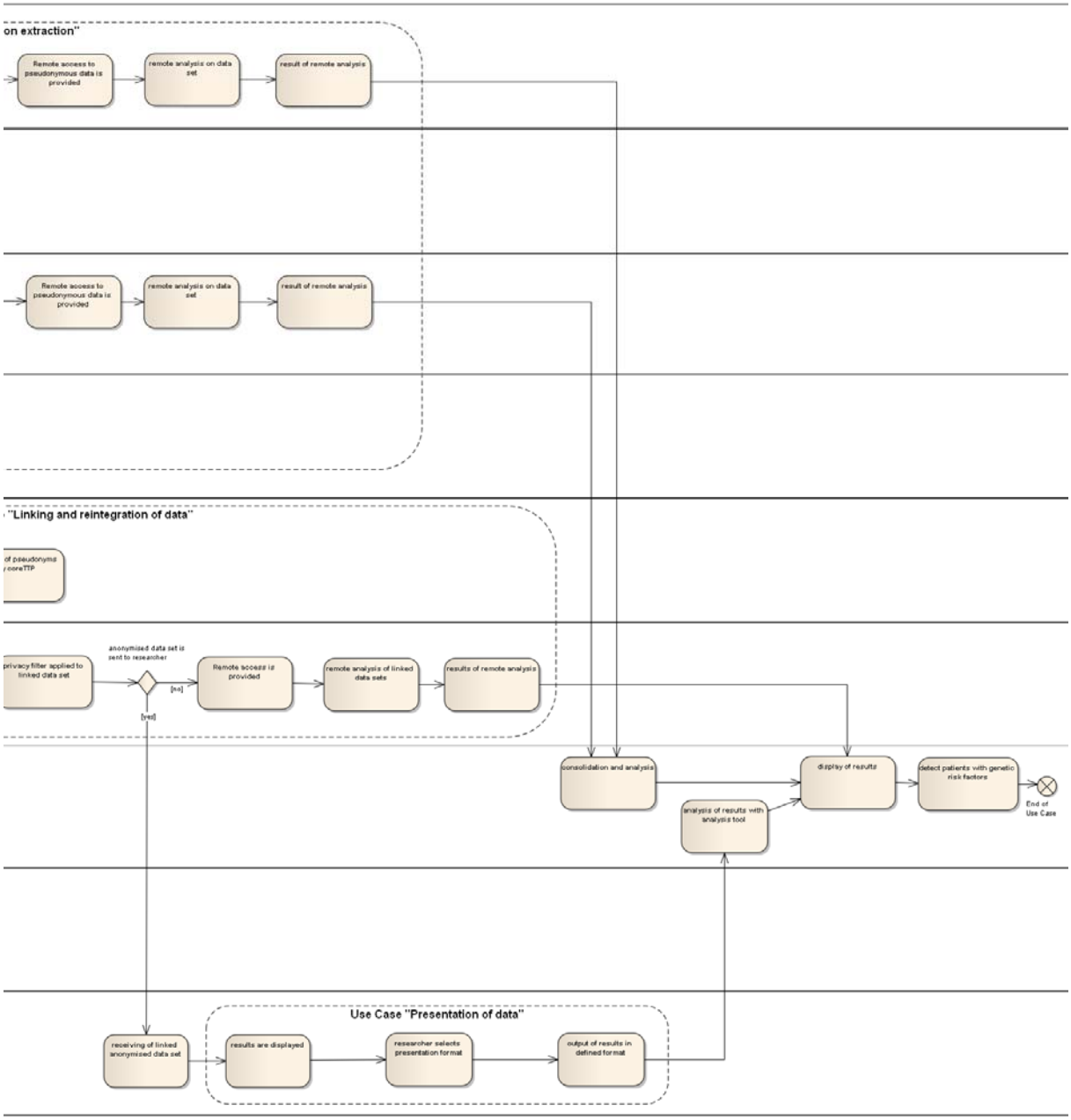
Fig. 21 Complete workflow of the Diabetes use case


Use Case "Authorisation of data selection, extraction and linkage
decision about access, data extraction and linkage
extraction is allowed
linkage is allowed
[yes]
[no]
stop
[no]
stop
decision about access, data extraction and linkage
extraction is allowed
linkage is allowed
[yes]
[no]
stop
[no]
stop
[yes]
Use Cas
[yes]
analysis of results / decision on selection and linking
researcher starts patient selection
[yes]
suitable quality is available / linking is possible
[no]
new search
selected data bases
Display results
definition of query and selection of data bases depended on quality analysis
list of selection criteria is provided

(continued)


Use Case "information extraction"
preparation for extraction, allocation of pseudonyms
extracted data is sent
[no]
Remote access to pseudonymous data is provided
initiation of selection
preparation for extraction, allocation of pseudonyms
extracted data is sent
[no]
Remote access to pseudonymous data is provided
[yes]
initiation of selection
[yes]
Use Case "Linking and reintegrati
TTP-1 receives pseudonym list
TTP-2 receives pseudonym list
Mapping of pseudonyms by coreTTP
mapping
[yes]
Use Case "T2D patient selection"
extraction and linking of data sets
privacy filter applied to linked data set
[yes]
selection of criteria and data bases and confirmation of query
list of selection criteria is provided

(continued)


on extraction"
Remote access to pseudonymous data is provided
remote analysis on data set
result of remote analysis
Remote access to pseudonymous data is provided
remote analysis on data set
result of remote analysis
"Linking and reintegration of data"
of pseudonyms coreTTP
privacy filter applied to linked data set
anonymised data set is sent to researcher
[no]
[yes]
Remote access is provided
remote analysis of linked data sets
results of remote analysis
consolidation and analysis
display of results
detect patients with genetic risk factors
End of Use Case
analysis of results with analysis tool
Use Case "Presentation of data"
receiving of linked anonymised data set
results are displayed
researcher selects presentation format
output of results in defined format

(continued)

Discussion of UML modeling of the Diabetes use case

UML based use case driven modeling is used for a static line of modeling covering class diagrams describing the concepts in the clinical research domain, and a dynamic line of modeling containing activity and sequence diagrams for the processes of the Diabetes and the gastro-esophageal reflux disease (GORD) use cases. After developing UML activity diagrams for all Diabetes sub-use cases, a single workflow description of the entire Diabetes use case was developed. The different roles of databases, data controllers and researcher were integrated with the concepts of the TRANSFoRm privacy framework (e.g. data controller, TTP, privacy filter). The initiator of the workflow is the researcher who defines the research question and executes the query. A special role of the GP/family doctor/non-interventional study using existing data was not considered because the use case deals with an observational study and direct involvement of the care area is in this case not necessary. The first step is the search for suitable database contents. Using the dynamic data discovery service with the Quality Tool, the researcher selects the databases and defines the search questions regarding the database metadata, e.g. availability of data, completeness of data, availability of policies, costs. These metadata is already available in a special repository, or has to be searched for in the databases. In addition, it might be only possible that the database make this data available on request. All three possibilities are included in the activity diagram. The results of the search / request are displayed, enabling the researcher to select suitable databases for the patient selection query. In particular, provided results will enable him to judge about the possibilities of record linkage.
According to the TRANSFoRm privacy framework [14] any access to primary care data, and any data extraction or data linkage must be authorized by the data controller of the database. Some databases provide search services for researchers and do only allow access to their data, but no extraction. A distributed approach (dataSHIELD [16]) has been suggested to overcome this limitation and to allow the pooling of searches of different databases. Indeed, large-scale data pooling has become important in epidemiology, particularly in the analysis of networks of studies, in public health research and in comparative international analysis. Such pooling not only supports large sample sizes but also reduces bias arising from access to restricted subsets of data. The dataSHIELD type of analysis is mathematically equivalent to a conventional “study level meta-analysis” but with all individual-level data remaining secure in the databases. Here the first step (estimation of regression coefficients and standard errors) is controlled remotely by the analysis centre of dataSHIELD and not carried out by the investigators at each study independently. The TRANSFoRm privacy framework [14] classifies genomic data as potentially identifiable data in need for additional protection. For linked data sets that contain genetic information, a privacy filter (e.g. an aggregation step) must be included in the data flow before the anonymised data are sent to the Query tool and the researcher.

Special elements

eHR

The entire workflow happens in the non-care zone and the research zone of the TRANSFoRm privacy framework [14]. It was not necessary to introduce the GP and the care zone into the workflow, because no necessity for patient contact or for access to patient data in the eHR was seen. All patient data is available in some form of care data register / database. In this context, it was necessary to consider that large care databases like NIVEL and GPRD have evolved to research service providers, employing their own data controller, ethics board and TTP. The workflow has incorporated this scenario in addition to the one for smaller databases that don't offer their own services.

TTP

For linkage of data different approaches are used in practice. Linkage can be performed at a central institution, such as Sweden Statistics with a linkage procedure based on a social number. Another approach would be to link within a non-care data base, as is planned for NIVEL [17]. In this case the data controller is clearly linked to the original database (NIVEL). Another approach would be to involve a coreTTP to bring together the TTPs of different databases. These different possibilities have to be taken into account in the modeling procedure. Only if this procedure is based upon national or regional regulations and policies, it can be implemented for practical use and will be applicable to different databases. This issue has been discussed in detail in the milestone on Trusted Third Parties and in the Confidentiality model (Deliverable 3.2). Here it is important to point out that according to complexity and lack of homogeneity of national rules and regulations on the European level, no single uniform solution can be offered. Instead, tailored solutions for linkage of specific data sources have to be worked out.

Dynamic Data Discovery Service

The Quality tool is part of the Dynamic Data Discovery Service, which was introduced in the activity diagram to collect system options and data quality information.

Distributed analysis

An important issue is whether data collected in one country can be transferred to another country for research. From the discussion so far, it became clear that full data transmission to the researcher is critical for many data providers, especially when data are linked with genomic data and the risk of re-identification is increased. In case linked data sources from different databases can be merged and transferred to the researcher, analysis is straightforward because all data are available for analysis in one place. If this cannot be achieved it has to be sorted out whether "distributed analysis" can be performed [16]. It has been shown convincingly, that for specific types of statistical techniques it is not necessary to analyse all data simultaneously (e.g, regression techniques). A possible approach would then be to start analysis in country A and to transfer the interim results to the researcher. In the

next step these interim results are brought together with the data in country B and analysis is updated in country B, and so on. If it is possible to conduct a distributed analysis, this strategy allows providing the statistical data needed for the study without data transfer between countries. In our case we have to explore whether case-control studies allow distributed data analysis.

8. Mapping of the Diabetes use case to PCROM and BRIDG

Use Case activity diagram

The following Diabetes use cases were modelled and incorporated into the complete clinical trials reference model: (1) present system options (e.g. preview counts / characteristics of available research data), (2) authorization of data extraction and linkage (e.g. conditions for the need for informed consent), (3) selection of patients with type 2 diabetes, (4) extract information (of selected patients), (5) reintegrate data from different databases at the level of individual patients, (6) presentation of data for further analysis. In the sub-use cases 3, 4 and 5 provenance requirements must be considered. For the T2D use case a detailed description of data extraction / remote access and data linkage by a TTP according to the TRANSFoRm privacy framework was developed. But many of these added elements, like remote access, data extraction, allocation of pseudonyms, pseudonym list, have more to do with the data model of TRANSFoRm and less with the information model.

Relationship and relevance of Diabetes Use Case objects for PCROM and BRIDG

Analysis of the Diabetes use case activity diagram generated a number of information objects that were analysed according to their relevance as class objects in PCROM and BRIDG (table 5): genetic database, care database, controller of database, TTP, eligible patient, inclusion and exclusion criteria, informed consent, patient, patient identifier (patient identifying information), prescription, research question, researcher, search, symptoms, recipient (for temporal data storage), metadata of database, data extraction, database access, linkage, patient selection, pseudonym, remote access, mapping of pseudonyms, remote analysis, privacy filter, genetic risk factor, Dynamic Data Discovery Service, Query Tool, and anonymised data. It was examined if a new information object is already included in a class or in a relation, and what its relation to existing classes may be. The notion -> refers to the affiliation of an object to the data model, rather than to the information model.

No.	Class / Domain Objects Use Case Diabetes	Comment / Description	PCROM objects	BRIDG objects
1	genetic database	-> data model (data source)		

2	care database	-> data model (data source)		
3	controller of database	-> data model (role)		
4	TTP	an organisation	Organization	
5	eligible patient	is person, is result of eligibility search = process activity	potential participant (singular)	
6	inclusion and exclusion criteria		EligibilityCriteria	DefinedEligibility Criterion, DefinedInclusion Criterion, DefinedExclusion Criterion
7	informed consent	activity (?), potential participant consents (=relation)	consented as relationship	StudySubject, StudyProtocol (not directly dispayed)
8	patient	is a person	patient	Subject
9	patient identifier (patient identifying informa-tion)	-> data model		
10	prescript-ion	process activity in medical care (prescription data base -> data model)	activity	Activity
11	research question		study purpose	StudyObjective
12	researcher	is a person	Investigator, StudyActor	QualifiedPerson, ResearchStaff, ResearchOrganiza tion
13	search	-> process activity (no information object)		

14	symptoms	result of an observational activity	observation activity result	DefinedObservation result
15	recipient (temporal data storage)	-> data model (role)		
16	metadata of database	-> data model		
17	data extraction	-> process activity		
18	database access	-> process activity		
19	linkage	-> process activity		
20	patient selection	-> process activity		
21	pseudonym	-> data model		
22	remote access	-> data model		
23	mapping of pseudo-nyms	-> data model		
24	remote analysis	-> process activity		
25	privacy filter	-> data model (software concept)		
26	genetic risk factor		AssessmentResult	
27	Dynamic Data Discovery Service	-> data model (software concept)		
28	Query Tool	-> data model (software concept)		
29	anony-mised data	-> data model		

Table 5: Relationship and relevance of Diabetes Use Case information objects for PCROM and BRIDG

Many of the analysed objects belong to the data model as data source or as software

concept. Other objects like prescription, eligible patient, inclusion and exclusion criteria, informed consent, patient, research question and researcher were already selected in the GORD use case. After screening for their relevance, following objects were left over. These objects seem therefore suitable to be included in PCROM (table 6).

No.	Object
1	TTP
2	prescription

Table 6: Objects of Diabetes use case to be included in PCROM

The modelling resulted in following changes in the PCROM model:

- Two objects (TTP and prescription) should be included in the information model.
- All relevant information objects are already in PCROM available (e.g. eligible patients, prescription, and researcher).
- TTP and prescription are added to PCROM (see also use case GORD for the other objects included in PCROM).
- The other objects are either data model concepts or a process activity with no connection to the information model.

9. Comparison of Information models

Structural differences of different information models

The clinical workflow model forms the basis of our information model. At first we searched for entire information models, or concepts and information objects that could be used for the TRANSFoRm use cases. There exist already a number of models dealing with the representation of information for clinical research. Developments in the following standard models were examined: openEHR, HL7 RIM, CTOM, CDISC SDM in view of the underpinning needs of the clinical models and actual clinical practice.

CTOM

The information model of The Clinical Trial Object Model, CTOM version 1.4 [18] is ruled by three classes: activity - observation – assessment. Activity is connected with procedures (surgery, imaging,..), observation is connected with histopathology, clinical results,… and assessment is connected with AE, diagnosis,… CTOM has been included in the initial domain modelling process of BRIDG [19, 20], though this information model is very much oriented towards cancer research. The subclasses of Activity contain the classes ActivityImpl (performing healthcare procedures on, or administering treatments to, a subject), ImagingImpl (method of tumour detection that involves obtaining a picture of the tumour, opposed to a biochemical test or direct observation (biopsy, endoscopy, etc.). ProcedureImpl (form of diagnostic, treatment or intervention); a healthcare activity experienced by a subject in a study. RadiationImpl (radiation treatment), SpecimenAcquisitionImpl (gathering and/or locating the sites for body samples, such as, urine, blood, biopsies, etc.), SubstanceAdministrationImpl (applying, dispensing or giving agents or medications to subjects), SurgeryImpl (surgery).

The subclasses of Observation in CTOM cover: ClinicalResult (results of a medical test that help determine a diagnosis, plan treatment, check to see if a treatment is working), Histopathology (microscopic study of characteristic tissue abnormalities), LesionDescription (process used to determine the extent of any localized or abnormal change in the structure of part of an organ or tissue.)

The Clinical Trials Object Data System (CTODS) was formerly known as CTOM. CTODS provides a collection of APIs and will provide a way for cancer research organizations to share or access de-identified data via the CTODS Viewer and CTODS API.

The information object Observation is connected to the objects Assessment and Activity by ObservationCollection. This is the only indication on the collection of clinical data. For example, Diagnosis goes not into Observation, but into Assessment.

PCROM

The main axis describes the information objects for participant - activity (observational activity, assessment activity, intervention activity) – study event. PCROM [3] is a good solution and an improvement compared with CTOM and BRIDG. Though, the position of timeline is difficult to understand. PCROM is an integral part in the work on the electronic Primary Care Research Network (ePCRN) [21], a National Institutes of Health-funded Roadmap Initiative project. The ePCRN is a secure, grid-based information system infrastructure that facilitates the conduct of randomized clinical trials in primary care. PCROM is published as class diagram with associated definitions that capture the components of a primary care RCT. Forty-five percent of PCROM objects were mapped to BRIDG, 37% differed in class and/or subclass assignment, and 18% did not map [3]. PCROM is organized into 3 interconnected sub-models: the trial process, which represents the information used by and/or generated by the individual steps or activities in a RCT. The primary class in this sub-model is the Activity such as administering an experimental treatment. Principal types of RCT activities are interventions or experimental treatments, assessments of patient status and condition, and observations of data results from measurement procedures. The Trial Information sub-model focuses on classes that describe the nature of the trial itself with the core concept being the Study class. For illustration the subclass of Intervention CognitiveIntervention is shown. There are differences between PCROM and BRIDG: with respect to interventions, BRIDG focuses on the objects of a regulated clinical trial such as a pharmaceutical entity, radiotherapy, or procedure. PCROM considers these to be subclasses of a PhysicalIntervention. The PCROM takes a much simpler view of the event flow within a study than does the BRIDG model. PCROM represents events as a timeline consisting of a collection of one or more activities occurring within a specified time in which where those periods are related in a chronological sequence.
PCROM provides no class for clinical data collection, but it connects Examination, Imaging and Interview with ObservationalActivity. The question is, if Interview can cover form based data collection, like it is done with a web questionnaire. We decided against this possibility.

BRIDG Model

The BRIDG Model [19] is an instance of a Domain Analysis Model (DAM). It was constructed to be independent from any concrete implementation. Thus, the semantics of BRIDG is restricted to the “problem domain”. In particular, a DAM contains no semantics that is based on a particular “solution space”. In the case of the BRIDG Model, the domain-of-interest is protocol-driven research and its associated regulatory artefacts (e.g. data, organization, resources, rules, and processes, focusing on pharmacological, physiological, or psychological effects of a drug, procedure, process, plus associated regulatory artefacts). BRIDG is represented using the Unified Modelling Language (UML). However, the model uses only the most obvious and easily understandable constructs of UML. Since the scope of BRIDG is very large, several sub-domain models were provided:

- Adverse Event Sub-Domain: the Adverse Event sub-domain covers safety related

activities, like detection, evaluation, follow-up and reporting.

- Common Sub-Domain: represents the semantics that are common to all (or most) of the other sub-domains and might even be common to any healthcare-related domain analysis model (including people, organizations, places and materials).
- Protocol Representation Sub-Domain: is intended for those involved in the planning and design of a research protocol. It covers the characteristics of a study and the definition and association of activities within the protocols (including “arms” and “epochs”).
- Regulatory Sub-Domain: is intended for those involved in the creation and review of submissions to regulatory authorities.
- Study Conduct Sub-Domain: is intended for those involved in the execution of a research study, covering requirements from those involved in clinical trials. It focuses on the activities of conducting the study as well as the results from those activities.

The goal of the BRIDG Project is to produce a shared view of the dynamic and static semantics of the domain of protocol-driven research and its associated regulatory artefacts. The BRIDG Model spans the lifecycle of a study from the planning of a study through the implementation, execution and evaluation of the study.

- Defined activities are the characterization of a kind of activity, i.e. they define “what” an activity is.
- Planned activities are the association of defined activities to a particular study. This association also includes the characterization of the timing of activities.
- Scheduled activities are the instantiation of a planned activity for each subject on a study; this is sometimes referred to as the “subject calendar”.
- Performed activities represent the execution of activities for actual subjects on a study and the results that come out of those activities. Because domain experts have identified a need to capture what was actually done to a subject, as opposed to what was intended to be done.

This classification is rather confusing, because the model creates redundancy, for example between study activities (defined activity, planned activity). For our modelling we concentrated on protocol representation and therefore considered only defined activity and planned activity, but not performed activities.

The new version 3 of BRIDG contains cross-referencing between BRIDG and HL7-RIM. For the representation of information objects this link was examined. The list “BRIDG UML to RIM mapping” provides a comprehensive comparison of BRIDG elements with the corresponding RIM-based components. The comparison shows that from BRIDG to RIM some research specific semantic is lost. For example, all three BRIDG elements PerformedClinicalInterpretation, PerformedClinicalResult and PerformedDiagnosis are represented in RIM by ObservationEventResult. A Performed Observation (fig. 22) consists of several ObservationEvents, but is also connected to SubstanceAdministrationEvent, ProcedureEvent, etc. In general, the RIM representation seems to be far away from the research process.

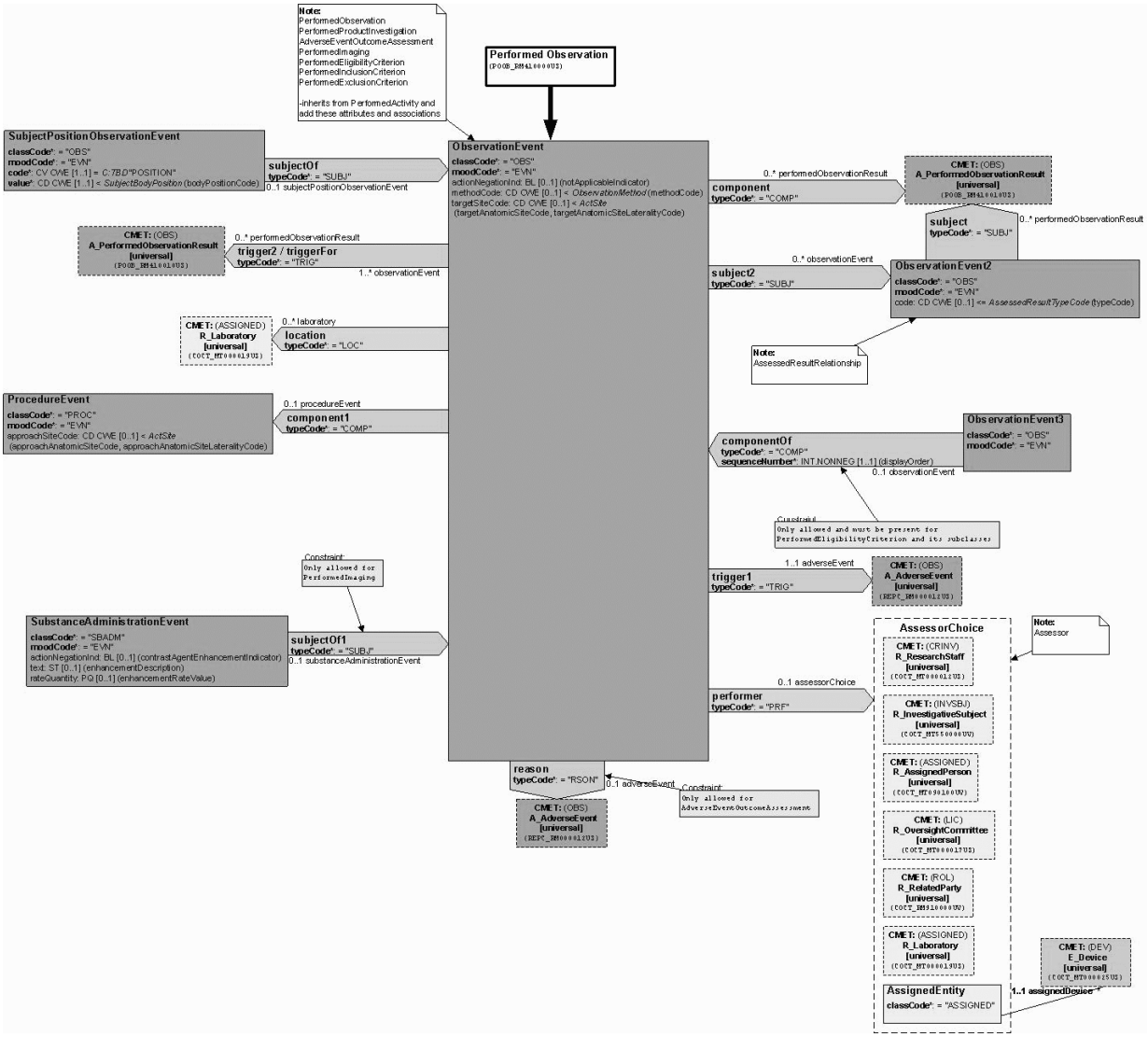


Fig. 22 RIM representation of Performed Observation (from: BRIDG version 3.03)

CDIDC SDM and CDASH

The Study Design Model [23] describes XML elements, their attributes, and their relationships that are relevant for study design. The terms elements and objects are used to reference the XML constructs of the same name, or alternatively, to reference objects and properties of those objects, respectively. The study design model is just a model of a study's design. It is therefore not a record of an individual study participant's way through the study. The study design model separates distinct aspects of study design into sub-components:

- Structure: Arms, Cells, Segments, Activities, etc.
- Workflow: progression through a study: decision points, branches, etc.
- Timing: time points of activities, usually relative to other structural elements of the study.

Study execution is outside the scope of this specification. However, the study design model was developed with the intent of ensuring that an execution runtime would have sufficient information to operate over a study design. The main axis of objects in SDM covers Study

Event – Activity – Segment. Structural elements are epoch, segments,... (their relation is somewhat difficult to understand, though). The Study Event is the main structural element of study design.

The CDASH [23] project identifies the basic data collection fields needed from a clinical, scientific and regulatory perspective to enable efficient and consistent data collection at the investigative sites. CDASH provides a set of essential/highly recommended eCRF fields. SDTM and CDASH are related. The SDTM and the SDTM Implementation Guide (SDTMIG) provide a standard for the submission of clinical trials data. Thus, CDASH is located early in the dataflow and defines a basic set of data collection fields that have to be present in the majority of CRFs. CDASH data collection fields were used as categories for the collection of clinical data in the information model. These safety domains are common to all therapeutic areas.

- Adverse Events (AE)
- Inclusion and Exclusion Criteria (IE)
- Comments (CO)
- Lab (LB)
- Concomitant Medications (CM)
- Medical History (MH)
- Demographics (DM)
- Physical Examination (PE)
- Disposition (DS)
- Protocol Deviations (DV)
- Drug Accountability (DA)
- Subject Characteristics (SC)
- ECG (EG)
- Substance Use (SU)
- Exposure (EX)
- Vital Signs (VS)

HL7 RIM

HL7 RIM (HL7 Reference Information Model) [24] is a generic information model developed by the Health Level Seven Consortium, designed to express the way data content, needed in a specific clinical or administrative context within the healthcare domain, can interoperate syntactically and semantically. It is part of the BRIDG package. RIM constituted the basis for developing the HL7 V3 messaging protocol specifications standard for healthcare interoperability.

RIM uses in principle four concepts:

- Entities: physical units such as things, human beings, organisations, groups
- Roles: time bound named functions for an entity, e.g. healthcare provider
- Acts: intentional actions, e.g. healthcare encounter, referral, intervention, etc.)

- Participations: links between an act and a role

The HL7 RIM is the most comprehensive information model for all aspects of the healthcare enterprise. In addition, the Clinical Document Architecture (CDA) was created as a document XML-based standard that specifies the structure and object that can include text, images, sounds, and other multimedia content for clinical documents structured recording and exchange.
Because of its breadth, the RIM tends to use generic class, attribute and association names that are not necessarily domain-friendly. In addition, the RIM is free of the constraints and business rules that apply to domain-specific models. Its aim is to provide a single set of reference semantics that can be leveraged across all healthcare domains. BRIDG semantics are not easily mappable to the HL7 RIM: the mappings are often not one-to-one, e.g. attribute to attribute. In fact, BRIDG class doesn't often map to a single RIM class (exceptions being concepts like "Person") and, likewise, a single attribute in the BRIDG model may map to a combination of RIM attributes or a collection of RIM data type properties.
In RIM for Adverse Event the ObservationEvent is connected via subjectOf4 with Document (universal). For the Performed Observation in RIM the ObservationEvent is connected through trigger2 with PerformedObservatioResult. On the other hand the SubstanceAdministrationEvent is connected via subjectOf1 with the ObservationEvent. These are quite complex relations and it is unclear what the relation between ObservationEvent and the collection of the observation could be.

EHR Information Model (openEHR Reference Model)

The EHR Information Model [25] is a model of an interoperable EHR from the ISO RM/ODP information viewpoint. It defines a logical EHR information architecture rather than just architecture for communication of EHR extracts or documents between EHR systems. The use of openEHR can be attributed to the formal acceptance of CEN 13606 [26] as a European and ISO standard. The aim of ISO EN13606 is to enable semantic interoperability in electronic health record communication. This standard is based on major aspects of the openEHR design approach. In principle, openEHR implementations can easily generate ISO EN 13606 communication extracts, because ISO EN 13606 represents specifications for the exchange of "EHR Extracts" [27]. Technically it may resemble a simplified version of the openEHR reference model. For example, the different ENTRY types of openEHR are mapped to a single ENTRY type in ISO EN 13606. Because ISO EN 13606 has not been implemented as a single standard but in a variety of locally customised forms, a revision has been proposed [28].
Of central importance in the openEHR information architecture is the EHR Class. The purpose of the EHR object is the root object and access point of an EHR for a subject of care. For data capture a Context Model of Recording is provided. This is necessary because, the openEHR model takes into account the importance of context in clinical activities, the real world is mapped to particular levels of the information model in a clear way, according the scheme in Fig.1. On the left side of the picture the context of data-entry is depicted, in which the information generated by a „healthcare event", containing „clinical statements", is added to

the EHR. A clinical statement is the minimal unit of information the clinician wants to record. Clinical statements have always temporal and spatial structure as well as data values.

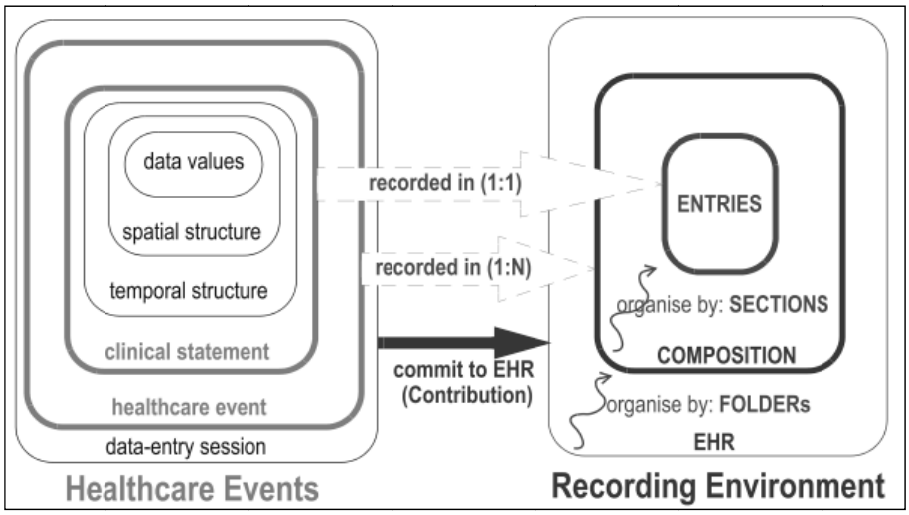


Fig. 23 Data capture context of openEHR, from [25]

All information created in the openEHR health record is expressed as an instance of a class in the entry package (fig. 23), containing the ENTRY class. An ENTRY instance is logically a single „clinical statement“, and may be a single short narrative phrase, but may also contain a significant amount of data, e.g. an entire microbiology result, an examination, a complex prescription. In terms of actual content, the Entry classes are the most important in the EHR Information Model, since they define the semantics of all the information in the record. It is obvious that this data context is different from the one employed in clinical trials.

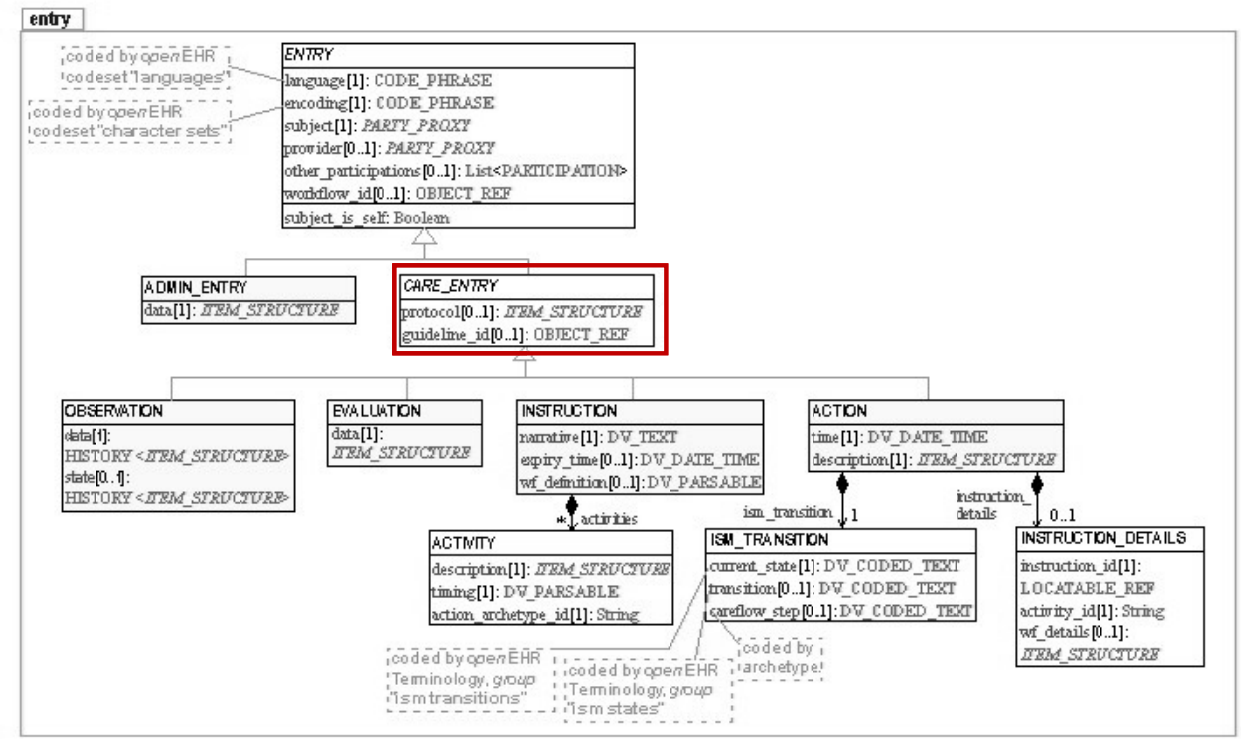


Fig. 24 Data entry unit of openEHR, “care entry” is highlighted, from [25]

Important attributes common to all Entry subtypes are subject: this attribute records the subject of the Entry as an instance of a subtype, and provider: the agent who provided the information (usually the patient or the clinician). A basic division occurs between clinical and non-clinical information. The CARE_ENTRY class is an abstract precursor of classes that express information of any clinical activity in the care for the patient, while ADMIN_ENTRY is used to capture administrative information (fig. 24).
One possible solution for our information model could be the incorporation of an ENTRY class into the PCROM analogous to the openEHR model. In the end, we choose another solution and introduced an overarching eCRF/Entry concept (see modelling).

aspect	CTOM Version 1.4 (2007)	PCROM Version 2.0	BRIDG (Version 3.02) Protocol Representation sub-domain	CDISC Study Design Model Version 1.0	comment
basic concepts	• activity • study event	• activity • study event	• study activity	• activity • study event	differences with respect to relation between activity and study event
representation of structural concepts of clinical trials	• study time point	• many	• epoch	• epoch • cell • segment	differences in granularity
link to study protocol	from activity via study participant via site to protocol	from study event via period via timeline to protocol	from study activity to sub-protocol version	protocol elements exist within categories: structure, workflow, timing	differences concepts (link via patient, standard elements of protocol)
represent-tation of care aspects	• health care site, health care participant	• health care site, patient	• health care provider (-subject)	----------	represented except for CDISC SDM

Table 7: Comparison between the key characteristics of information models

Relevance of the different models for our primary care research model

The usefulness of parts or the whole model for the representation of the TRANSFoRm information elements and the use case research processes was evaluated. For this purpose, several relevant information models were compared (table 7). All analyse information models, CTOM, PCROM, BRIDG, and SDM, all are mainly concerned with the clinical research domain. The concepts of study event, activity, study activity, etc. play important roles. Concepts from the overlapping area between care and research, like web questionnaire, invitation of patient, prescription, appointment (incl. GP visit), patient screening, consultation, etc. find no corresponding objects in these models. On the other hand, care related models, like openEHR, represent the care domain in such a complex way that they may be not useful for a research information model. Nonetheless, they contain single concepts, for example the “entry unit” of openEHR that can inspire the use of similar concepts for the research domain. From all models examined, the PCROM turned out to be the most suitable for the modelling of TRANSFoRm research requirements, though an extension and adaption is needed.

10. Development of an extended PCROM

From PCROM (Primary Care Research Object Model) to CRIM

The evaluation of different information models resulted in the conclusion, to use the PCROM and extend / modify it according to the requirements of TRANSFoRm. The fact that PCROM represents a proper basis for our modeling is not surprising, because the initial scope of PCROM was already the performance of randomized clinical trials and contains many objects that are relevant for clinical trials (table 8). It was validated by domain experts and underwent a comparison with the BRIDG model. Therefore, a considerable overlap with TRANSFoRm research processes exists. In addition, similar to our efforts to describe the TRANSFoRm use cases, the PCROM team used activity models and use cases to develop a class model consisting of objects and their static associations. PCROM domain objects cover already a large number of objects listed in the use case description and the DAM for GCP trials (table 8).

Activity	HealthCareSite	Protocol
ActivityActivityRelationship	Imaging	PtInfoModel
AdmministativeContact	Intervention	Rule
AdverseEvent	InterventionActivity	SiteCoordinator
AEReport	Interview	Specimen
AETreatment	Investigator	SpecimenCollection
Allocation	Laboratory	Sponsor
AssessementActivity	LabTesting	Statistician
AssessmentMethod	ObservationActivity	Study
AssessmentResult	ObservationResult	StudyActor
CognitiveIntervention	OperationalStandardsMonitoring	StudyAnalysis
ContactInformation	Organization	StudyCoordinator
DiagnosticTest	OversightCommittee	StudyEvent
EligibilityCriteria	Participant	StudyGroup
EnvironmentModel	Patient	StudyOutcome
EthicsApproval	Period	StudySite
EthicsReviewGroup	Person	SystemChangeIntervention
Examination	PhysicalIntervention	SystemService
FundingBody	PotentialParticipant	Timeline
		TrialRegistration

Table 8: PCROM 3 domain objects

All together 12 new information objects were added to extend PCROM. To indicate that a set of objects belong to the patient / GP (family doctor) relationship a new “Care related area”

was introduced (fig. 25). This care related area contains the GP (family doctor) with his main activities: appointment, consultation, encounter and prescription. The concept of encounter was introduced to have a high-ranking concept that can include the different forms of interactions between patient and GP. Encounter describes the professional interchange between a patient and a GP (family doctor) that is related to the episode of care. One or more episodes of care may be dealt with at an encounter. The concept of episode of care describes a health problem from its first presentation by the patient to the GP until the completion of the last encounter. It encompasses all contact elements related to that health problem. The GP can also receive a study relevant reminder for example a reminder to invite a patient to complete a PRO questionnaire.

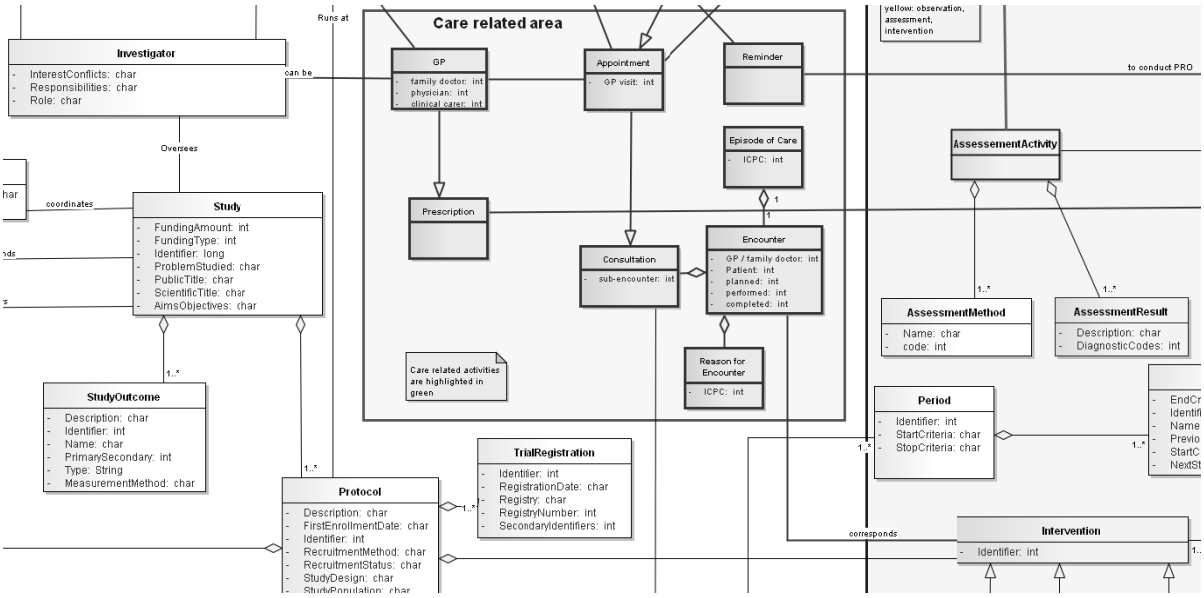


Fig. 25 The "care area" of the extended PCROM (framed) in relation to the neighbouring concepts of Investigator, Study and Protocol.

The GP (family doctor) is a person that does connect through the appointment with the patient. Therefore, appointment and AE become related to the research area. For example, an appointment because of an adverse event is an unscheduled study event. The step from the potential participant to the study participant is connected to "Invitation of patient". In a second step consultation was connected with encounter and "Episode of care". An encounter has always a reason for encounter and is part of an episode of care. An appointment of the patient with the GP can be a consultation (which is an encounter with a "Reason for Encounter") and connected to EligibilityCriteria outside the care area. Using these organising concepts from ICPC, our information model increases terminological clarity in this complex area of overlap between care and research.

A second area eCRF (ENTRY) was introduced as a high-level concept including observation and assessment concepts (fig. 26). The rational for this area lies in the complexity of the data collection process in clinical trials containing besides the actual data collection, aspects of

observations, assessment, intervention, study activity, examination and study event. The ENTRY area collects these aspects and combines them on a higher level. The web questionnaire as QoL assessment was connected to PatientReport; thus the WebQuestionnaire becomes an instance of PatientReport.
The PCROM was extended by 12 information objects with additional 3 episode of care related objects and two areas (fig. 27 and 28).

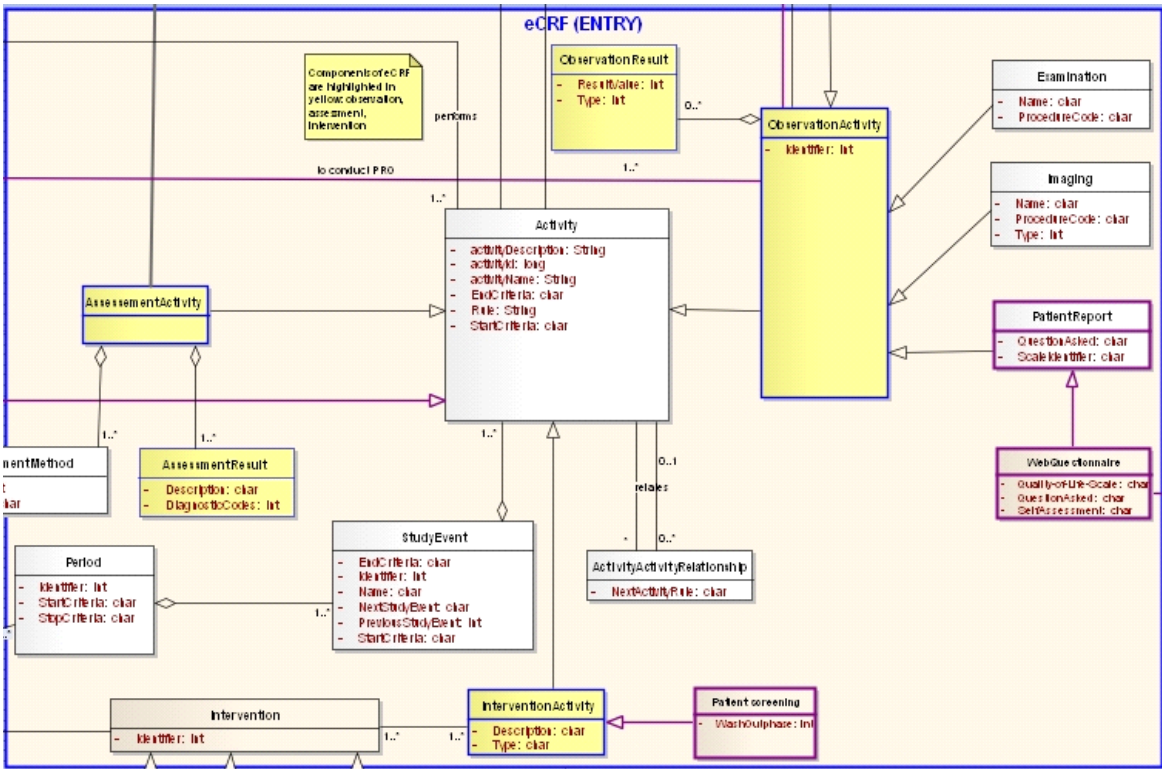


Fig. 26 "eCRF area" of the extended PCROM (objects relevant for data collection are highlighted in yellow, concepts added to PCROM are framed in violet)

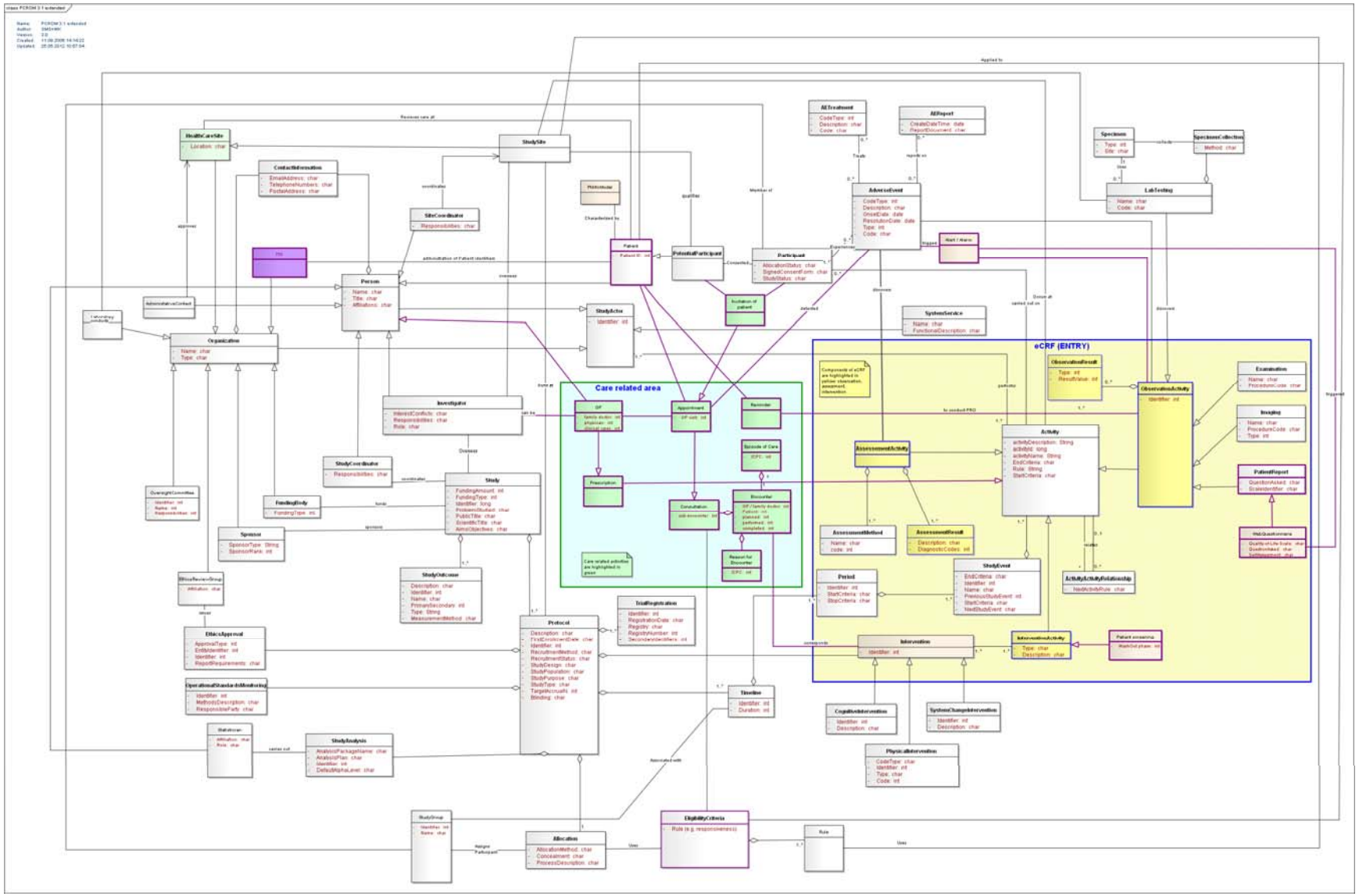


Fig. 27 Class diagram of CRIM (extended PCROM). Green: care related concepts, yellow: ENTRY area for data collection, violet frame: newly added objects

Diagram of the Clinical Research Information Model (CRIM)

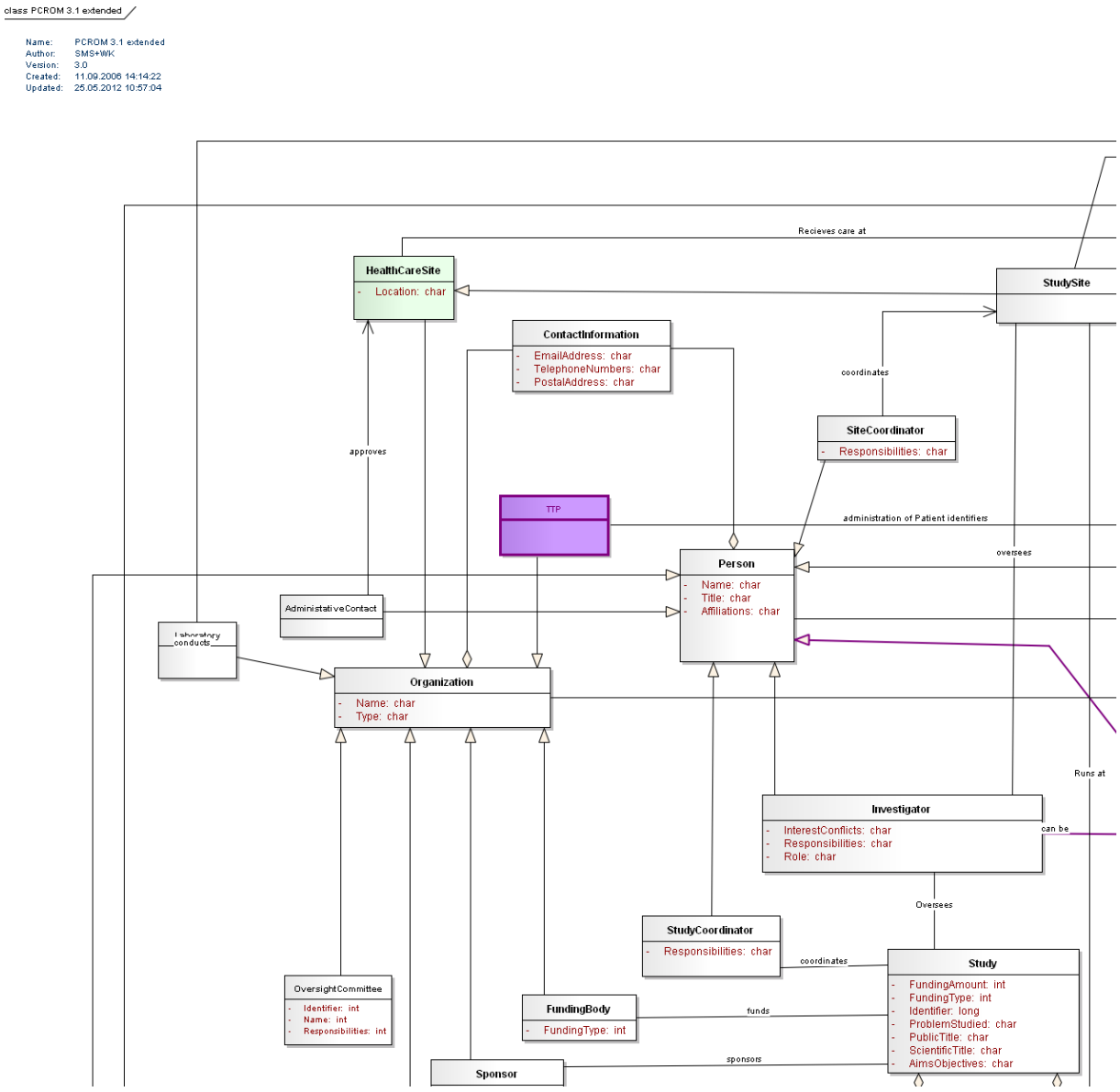
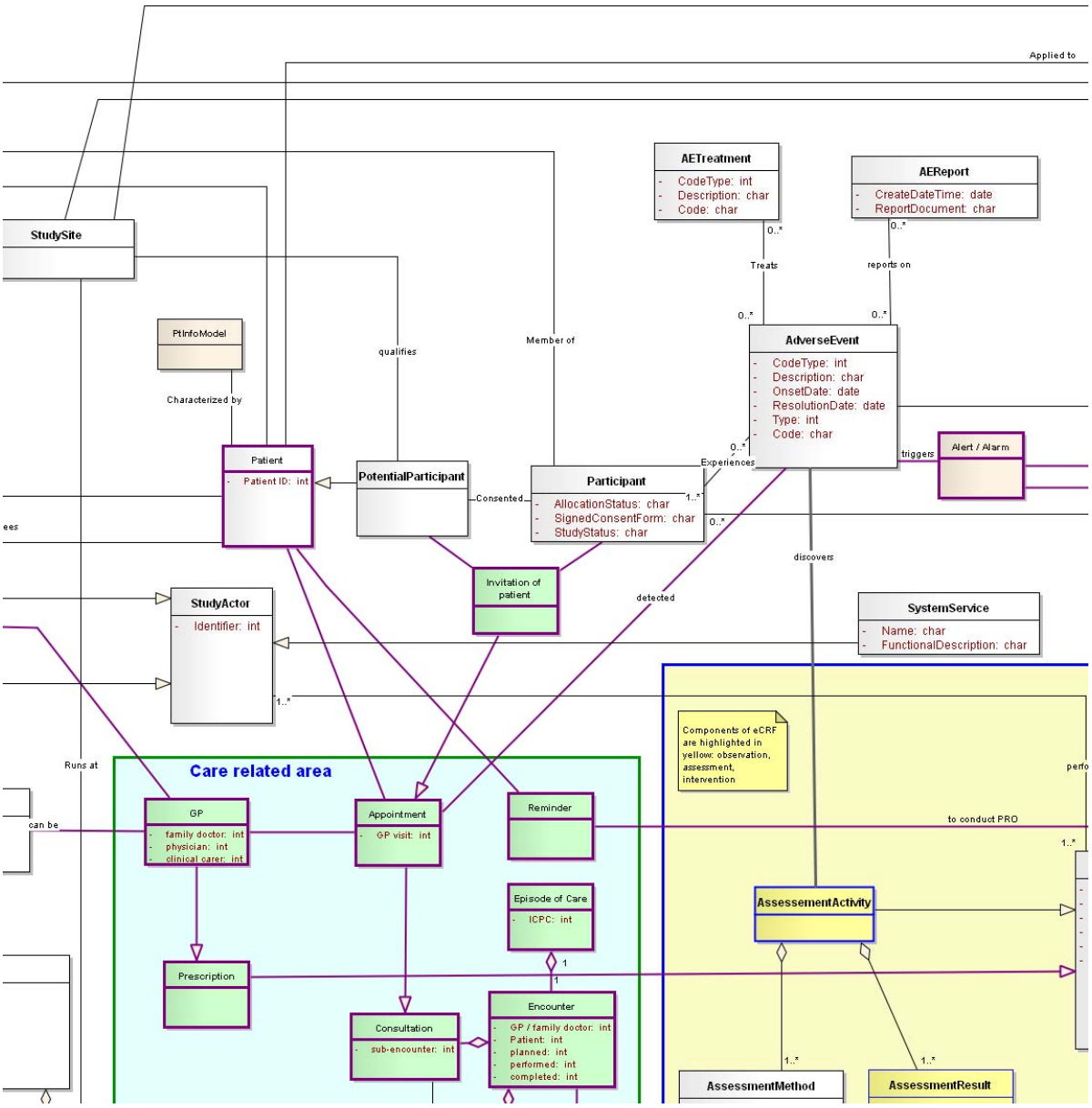
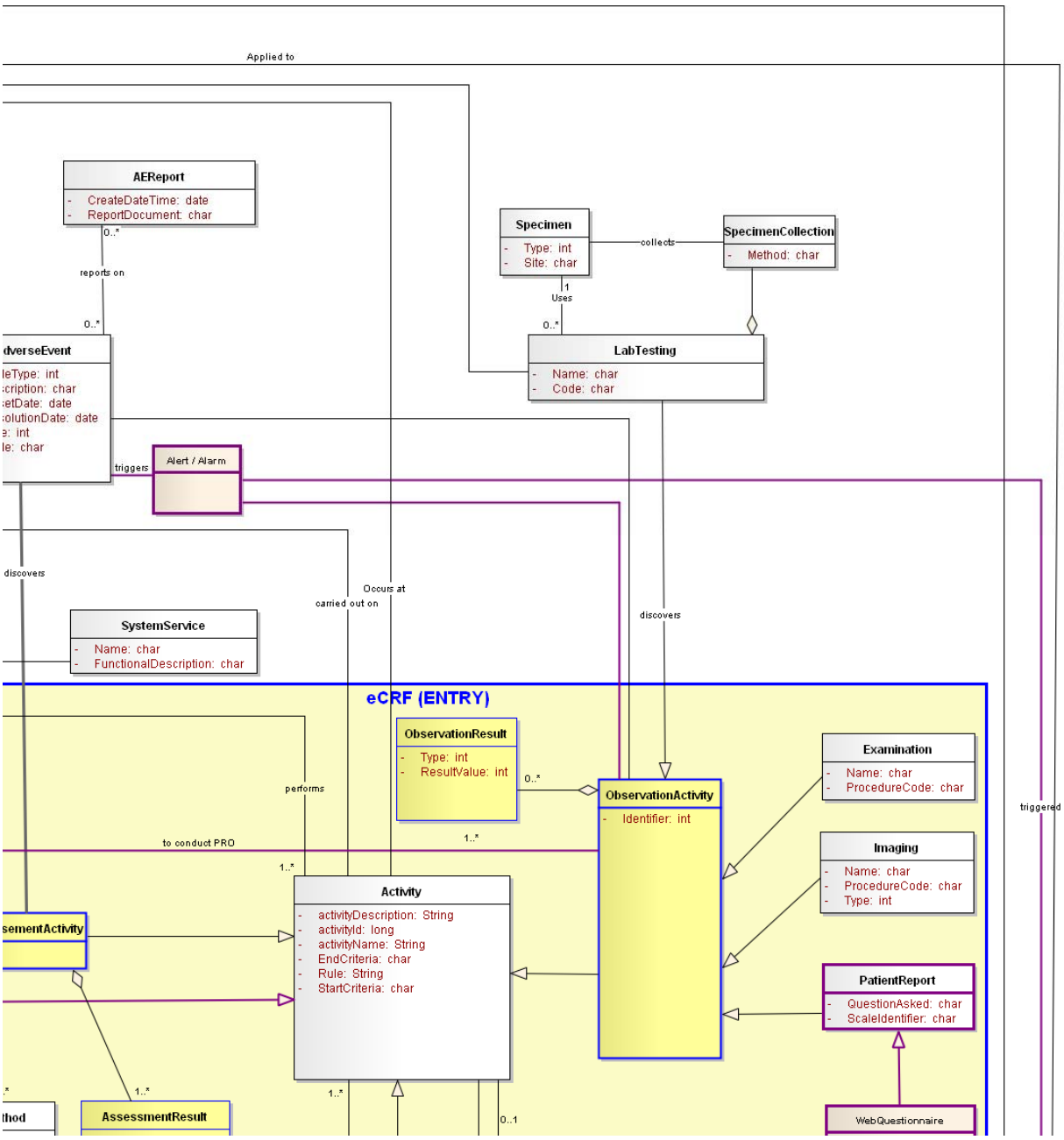


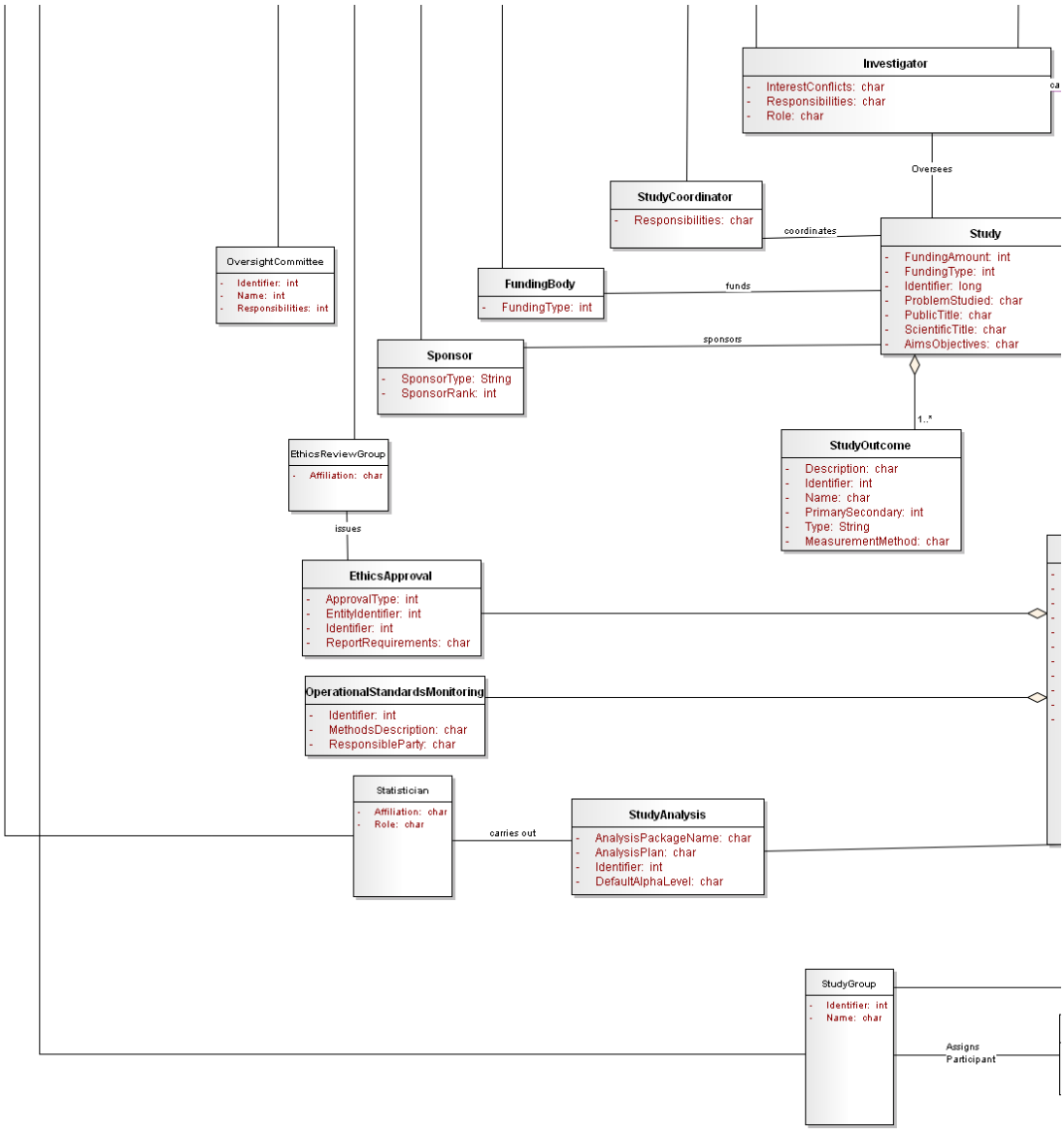
Fig. 28 Extended PRCOM – UML diagram, draft version 3 (distributed diagram)


Applied to
AETreatment
- CodeType: int
- Description: char
- Code: char
AEReport
- CreateDateTime: date
- ReportDocument: char
StudySite
Treats
reports on
0..*
PtInfoModel
qualifies
Member of
AdverseEvent
- CodeType: int
- Description: char
- OnsetDate: date
- ResolutionDate: date
- Type: int
- Code: char
Characterized by
Patient
- Patient ID: int
PotentialParticipant
Consented
Participant
- AllocationStatus: char
- SignedConsentForm: char
- StudyStatus: char
Experiences
triggers
Alert / Alarm
discovers
Invitation of patient
detected
StudyActor
- Identifier: int
SystemService
- Name: char
- FunctionalDescription: char
Runs at
Care related area
GP
- family doctor: int
- physician: int
- clinical carer: int
Appointment
- GP visit: int
Reminder
Components of eCRF are highlighted in yellow: observation, assessment, intervention
to conduct PRO
Episode of Care
- ICPC: int
AssessementActivity
Prescription
Consultation
- sub-encounter: int
Encounter
- GP / family doctor: int
- Patient: int
- planned: int
- performed: int
- completed: int
AssessmentMethod
AssessmentResult

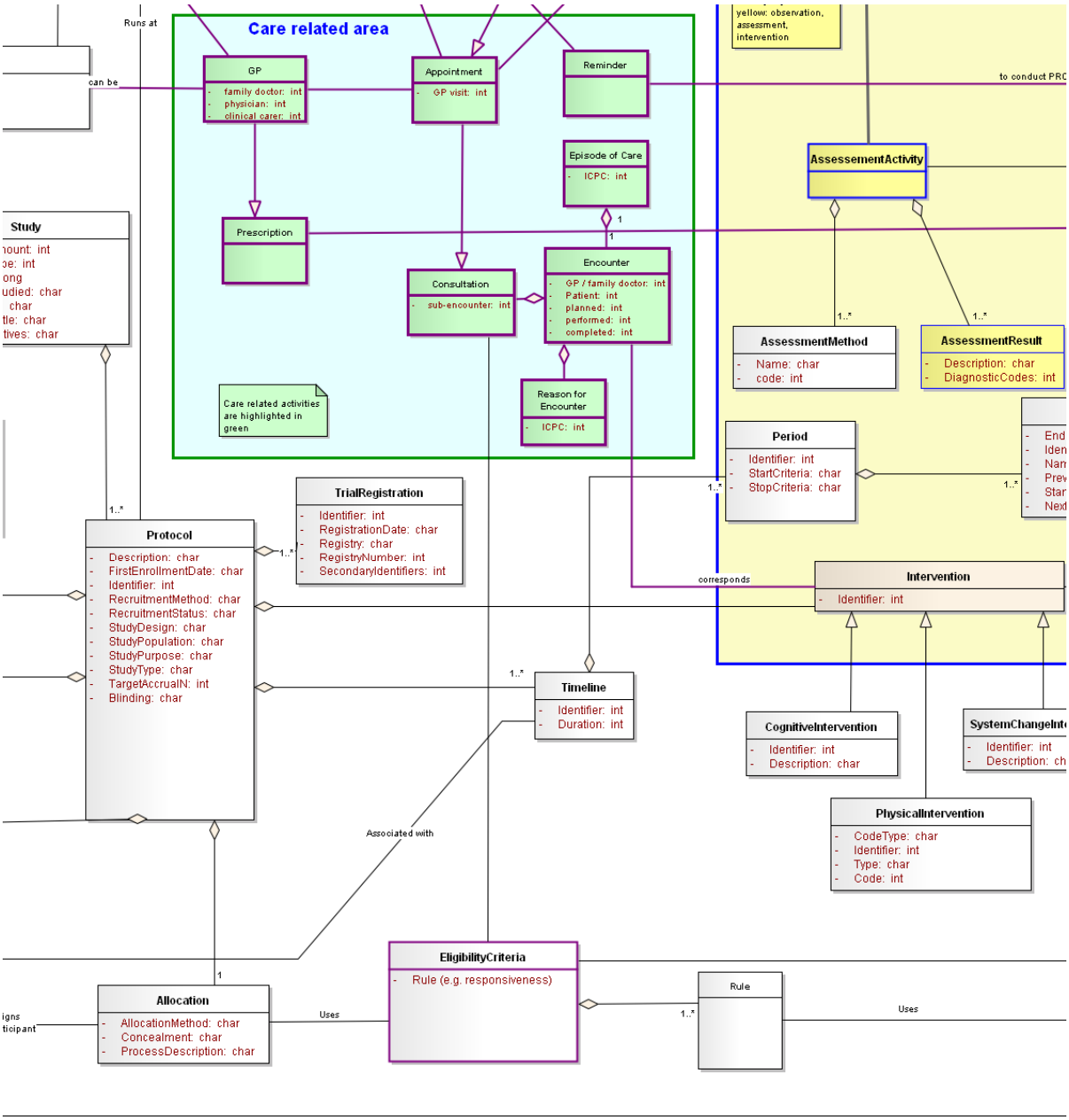
(continued)


Applied to
AEReport
- CreateDateTime: date
- ReportDocument: char
0..*
reports on
0..*
dverseEvent
leType: int
cription: char
etDate: date
olutionDate: date
e: int
le: char
triggers
Alert / Alarm
Specimen
- Type: int
- Site: char
collects
SpecimenCollection
- Method: char
1
Uses
0..*
LabTesting
- Name: char
- Code: char
discovers
Occurs at
carried out on
SystemService
- Name: char
- FunctionalDescription: char
discovers
eCRF (ENTRY)
ObservationResult
- Type: int
- ResultValue: int
0..*
1..*
performs
to conduct PRO
1..*
ObservationActivity
- Identifier: int
Examination
- Name: char
- ProcedureCode: char
Imaging
- Name: char
- ProcedureCode: char
- Type: int
PatientReport
- QuestionAsked: char
- ScaleIdentifier: char
WebQuestionnaire
triggered
Activity
- activityDescription: String
- activityId: long
- activityName: String
- EndCriteria: char
- Rule: String
- StartCriteria: char
sementActivity
1..*
1..*
AssessmentResult
0..1
thod

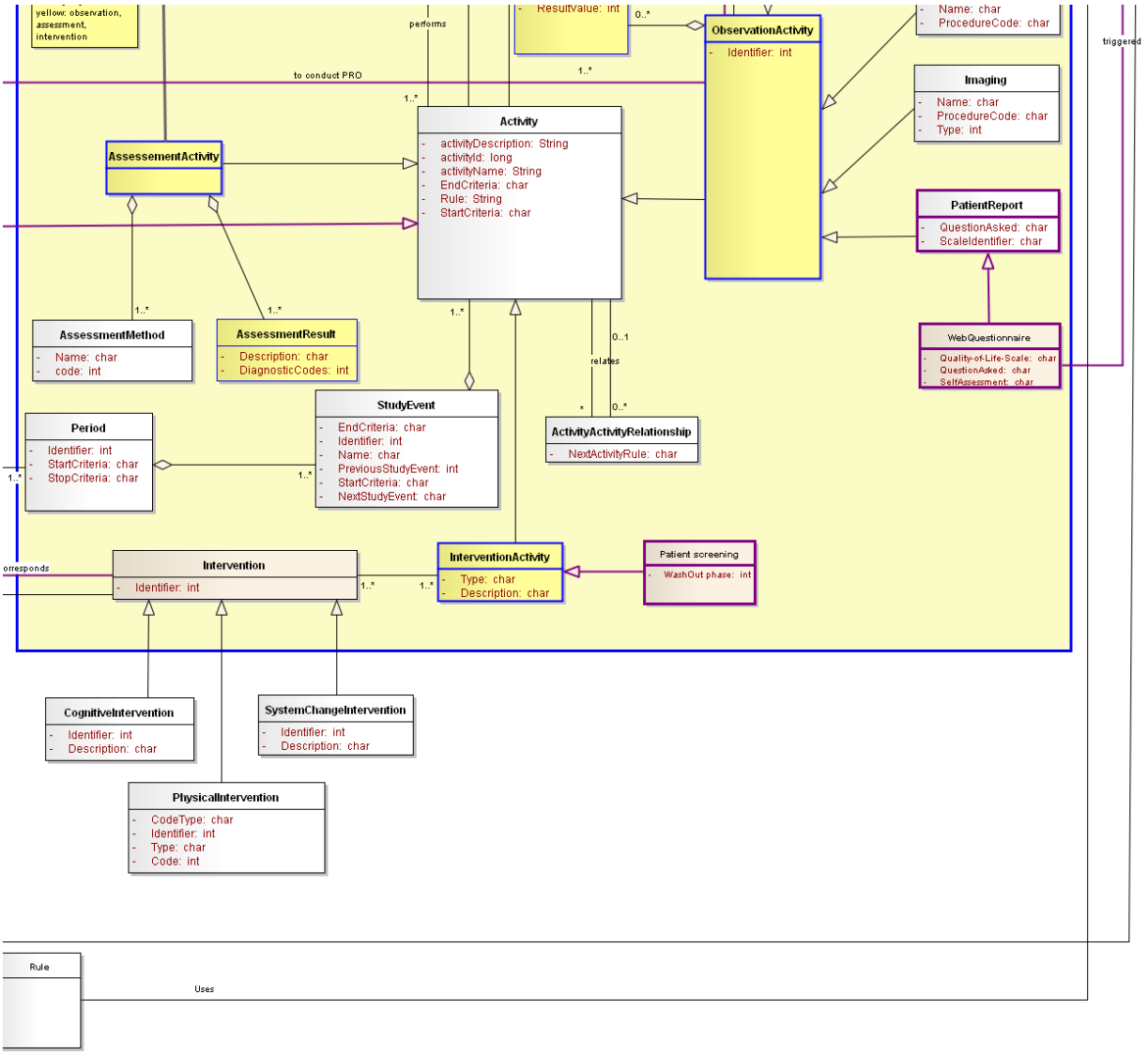
(continued)


Investigator
- InterestConflicts: char
- Responsibilities: char
- Role: char
Oversees
StudyCoordinator
- Responsibilities: char
coordinates
Study
- FundingAmount: int
- FundingType: int
- Identifier: long
- ProblemStudied: char
- PublicTitle: char
- ScientificTitle: char
- AimsObjectives: char
OversightCommittee
- Identifier: int
- Name: int
- Responsibilities: int
FundingBody
- FundingType: int
funds
Sponsor
- SponsorType: String
- SponsorRank: int
sponsors
1..*
StudyOutcome
- Description: char
- Identifier: int
- Name: char
- PrimarySecondary: int
- Type: String
- MeasurementMethod: char
EthicsReviewGroup
- Affiliation: char
issues
EthicsApproval
- ApprovalType: int
- EntityIdentifier: int
- Identifier: int
- ReportRequirements: char
OperationalStandardsMonitoring
- Identifier: int
- MethodsDescription: char
- ResponsibleParty: char
Statistician
- Affiliation: char
- Role: char
carries out
StudyAnalysis
- AnalysisPackageName: char
- AnalysisPlan: char
- Identifier: int
- DefaultAlphaLevel: char
StudyGroup
- Identifier: int
- Name: char
Assigns Participant

(continued)


Care related area
GP
- family doctor: int
- physician: int
- clinical carer: int
Appointment
- GP visit: int
Reminder
Episode of Care
- ICPC: int
Prescription
Consultation
- sub-encounter: int
Encounter
- GP / family doctor: int
- Patient: int
- planned: int
- performed: int
- completed: int
Reason for Encounter
- ICPC: int
Care related activities are highlighted in green
AssessementActivity
AssessmentMethod
- Name: char
- code: int
AssessmentResult
- Description: char
- DiagnosticCodes: int
Period
- Identifier: int
- StartCriteria: char
- StopCriteria: char
Study
Protocol
- Description: char
- FirstEnrollmentDate: char
- Identifier: int
- RecruitmentMethod: char
- RecruitmentStatus: char
- StudyDesign: char
- StudyPopulation: char
- StudyPurpose: char
- StudyType: char
- TargetAccrualN: int
- Blinding: char
TrialRegistration
- Identifier: int
- RegistrationDate: char
- Registry: char
- RegistryNumber: int
- SecondaryIdentifiers: int
Intervention
- Identifier: int
corresponds
Timeline
- Identifier: int
- Duration: int
CognitiveIntervention
- Identifier: int
- Description: char
PhysicalIntervention
- CodeType: char
- Identifier: int
- Type: char
- Code: int
Associated with
EligibilityCriteria
- Rule (e.g. responsiveness)
Allocation
- AllocationMethod: char
- Concealment: char
- ProcessDescription: char
Uses
Rule
Uses
Runs at
can be
1..*
1

(continued)


yellow: observation, assessment, intervention
performs
ResultValue: int
0..*
1..*
ObservationActivity
- Identifier: int
- Name: char
- ProcedureCode: char
triggered
to conduct PRO
Imaging
- Name: char
- ProcedureCode: char
- Type: int
Activity
- activityDescription: String
- activityId: long
- activityName: String
- EndCriteria: char
- Rule: String
- StartCriteria: char
AssessementActivity
PatientReport
- QuestionAsked: char
- ScaleIdentifier: char
AssessmentMethod
- Name: char
- code: int
AssessmentResult
- Description: char
- DiagnosticCodes: int
WebQuestionnaire
- Quality-of-Life-Scale: char
- QuestionAsked: char
- SelfAssessment: char
relates
0..1
0..*
StudyEvent
- EndCriteria: char
- Identifier: int
- Name: char
- PreviousStudyEvent: int
- StartCriteria: char
- NextStudyEvent: char
Period
- Identifier: int
- StartCriteria: char
- StopCriteria: char
ActivityActivityRelationship
- NextActivityRule: char
orresponds
Intervention
- Identifier: int
InterventionActivity
- Type: char
- Description: char
Patient screening
- WashOut phase: int
CognitiveIntervention
- Identifier: int
- Description: char
SystemChangeIntervention
- Identifier: int
- Description: char
PhysicalIntervention
- CodeType: char
- Identifier: int
- Type: char
- Code: int
Rule
Uses

(continued)

11. Eligibility Criteria Information Model (ECM)

Introduction

Because querying of databases plays an important role in TRANSFoRm, the information model was extended by an eligibility criteria model (ECM). This extension will make the information model more applicable for TRANSFoRm search applications, like the Query Tool. Thus, an eligibility criteria model was developed as part of CRIM that was designed to describe a generic way to represent inclusion and exclusion criteria defined in clinical trials. To guarantee flexibility in realization, the ECM) was integrated in a “black box” manner into CRIM as an added module. Therefore, the extended PCROM and ECM are presented as separate diagrams, though they belong to the same CRIM.

As starting point for the development of ECM not a specific eligibility language or query framework was used, but common inclusion / exclusion criteria were analysed and based on this analysis a generic and simple model was constructed. This approach was similar to one used by Austin et.al. [29], but the conclusions drawn from the analysis of eligibility criteria differ. The main difference is that we do not use a system of different anchors (e.g. age anchor, event anchor and the concept of constraints (e.g. string constraint, data constraint). For the development of the ECM new concepts (like filter, operator) were introduced and adapted to the needs of TRANSFoRm. Analysis of eligibility criteria shows that temporal criteria play an important role in the definition of inclusion / exclusion criteria and therefore, the implementation of temporal relations has to be included in the development of ECM. But the representation of temporal relationships for data querying can become complicated [30]. In about one third of clinical trials examined, the timing of clinical assessments or interventions was well defined. For a query tool to be useful for TRANSFoRm the conclusion can be drawn that specification of temporal relations must be included and has to be as simple as possible. In principle, there are two approaches to deal with temporal relations in queries: first, creation of a sophisticated internal representation of time in the database with an equally sophisticated temporal query language and second, use of a simple, time stamped database combined with a sophisticated temporal query language [30]. Because in TRANSFoRm standard database representations will be queried the latter approach seems to be useful.

The information model of eligibility criteria should support the representation of inclusion and exclusion criteria with temporal constraints used in most protocols [31]. The ECM can serve as a starting point to understand the way inclusion and exclusion criteria are established and designed to be consistent with CRIM and CDIM. To make querying as simple as possible, it is advisable to build a query concept that is easy, extensible and adaptable to all possible types of clinical concepts, coding and categorizations. In this sense, the core part of the ECM consists of a generic model that can in principle be used for different domains (e.g. querying of biobanks). In this context the developed ECM is flexible enough to be used together with linear temporal logic, temporal SQL or some Temporal Knowledge

Representation.

Main Concepts of the Eligibility Criteria Model

Single Criterion, Rules and Operators

The basic building block of the ECM is the SingleCriterion (fig. 29). The SingleCriterion is defined by parameters (e.g. Time, Code, Category). Inclusion / exclusion criteria can be composed of one or more of these units and are expressed as Rules. Each single criterion can be combined with either temporal or logical operators. Thus, the Operator relates the SingleCriterion and the Criterion Group (Rules) with each other. The Operator plays an important role in ECM, because it is a generic concept to enable Boolean and temporal relations. The TemporalOperator combines SingleCriteria by a temporal operator employing the following functions: AFTER t; WITHIN t and BEFORE t (t = time span in years, months, days, hours or seconds). The BooleanOperator employs Boolean combinations within Rules, using combination of “AND”, ”OR”, “AND NOT” and “OR NOT”.
In addition, for query definition two types of processors can be applied: Comparator and Filter (fig. 29). Time can be represented as an event or as an interval (parameter). For the querying of events (e.g. parameters/clinical characteristics) the definition of temporal filters is useful to be able to index events to which other events are related on a time scale. These filters should cover the identification of the earliest, most recent, any event or all events. In addition, an age filter was defined considering the central role age plays in EHR data. The age filter uses a criterion which is derived from the difference between the date at which a clinical characteristic was recorded and the date of the patient’s birth.
In general, all eligibility criteria, their groups and the single criterion, can have an associated value of True, False, and Unknown. This statement has been expressed as attributes in the ECM. Though, most eligibility criteria in the query process are true or false, a search for unknown eligibility criteria seems to be impractical. Because, an exclusion criterion can be seen simply as the negation of an inclusion criterion, in the model we have no distinction between inclusion and exclusion criteria. In the class "Rules" inclusion and exclusion rules were added as attributes. Rules are either compositions of logically combined SingleCriterions or temporally combined SingleCriterions. In principle, Rules can be applied recursively to themselves to construct more complex units. There exist some limitations in EMC: in a single rules group the usage is limited to one operator type (temporal or logical (= Boolean) operators and the usage of temporal operators is limited to SingleCriterion.

Temporal operators

A precondition to the implementation of time operators is the use of a time stamped database with storage of the valid time of an instant event. The following operators are needed to describe all relevant temporal relations:

- before x (e.g. years, months or days)
- after x (e.g. years, months or days)

- within x (e.g. years, months or days)

These are fewer relations than is covered by several query models that are capable of handling “Allen operators”. But our analysis of inclusion / exclusion criteria showed that the implementation of a full set of Allen operators may not be necessary for most queries in the TRANSFoRm context. This, however, has to be evaluated in depth. Though, we have implemented the simplification that time operators can only link two events, this may be an unnecessary restriction. Because we want to design the ECM as simple as possible, we suggest using this limitation as starting point for a first iteration of improvements.

Boolean combinations

Queries are normally defined according to the inclusion and exclusion criteria defined in the study protocol. For the formulation of queries, each individual criterion (inclusion or exclusion) is transformed into a separate query statement. This query statement corresponds in the model to Rules. There should be no substantial restrictions on the definition of Boolean combinations within Rules, any combination of “and”, ”or”, “and not” and “or not” should be possible. It is not necessary to enter queries as a conjunctive (disjunctive) normal form but as it is defined in the text of the criterion. During building the query, all individual inclusion criteria will be linked with “and” and all exclusion criteria with “or”. For inclusion of a patient in a study, all of the inclusion criteria must be fulfilled. On the other hand, patients are not eligible for a study, if they fulfill one or more exclusion criteria. In this way, inclusion and exclusion criteria are logically combined.

Combination of events using temporal and Boolean operators

In the ECM it is possible to combine single events or a group of logically combined events by a time operator. There exist no restrictions of combining events/groups of events combined by a time operator with other events or group of events by Boolean operators. Events or event groups that are already linked by a time operator cannot be linked by a time operator with any other event or event group.

Points to consider: dependencies and comparisons

In ECM, Operator is either a Boolean or a temporal operator. SingleCriteria can be combined by a temporal operator. This requires some temporal relations between events (e.g. a time-stamped database). The following operators are needed as a minimum: AFTER t, WITHIN t and BEFORE t, where t is a value to define a time span. In contrast to the Operator, the Comparator is a concept to define a comparison between a value of a parameter, e.g. lab test HbA1c = 9,1%, and a specified cut-off value, e.g. HbA1c > 10 %, in which the result of the comparison can be true or false. We distinguish between a NumericComparator (comparing numeric values, with following functions: <,≤,=,≥,> and a StringComparator (comparing strings values, such as diagnosis = DT2, with the following functions: “equals”, “like” and “contains”.

This more generic concept was proposed to be able to extend if necessary the information model with new comparators, such for example by a HierarchyComparator that would examine if the value of the parameter (e.g. clinical concept in ICD 10) is higher or lower in a defined hierarchy.
ECM proposes a two-step selection process. Filter operates only on a ResultSet (it separates values) (fig. 30). A QuantityFilter is used on a SingleCriterion to screen for e.g. all patients that have at least three records of lab values HbA1c > 10%. In contrast, AgeFilter provides for the specification of the difference between the date a clinical characteristic was recorded and the age of the patient at that time. For example, it screens for patients that had been diagnosed with DT2 before the age of 18. Here the “EVENTAGE < t” for the SingleCriterion DT2, where t is a value to define a time in years. EVENTAGE = the age of the patient when the EVENT occurred. In addition, TemporalFilter is employed to screen for events (e.g. parameters / clinical characteristics). With this filter an event can become an index event to which other events are related on a time scale. Thus, TemporalFilter should cover the identification of the MOST RECENT (FIRST), LAST (LATEST/EARLIEST), ANY or ALL events as index elements.
The idea of the filter is to allow researchers to operate on / post-process a result set (value list) after a query has been launched and a result is returned (fig. 30). Using Filter, the returned result set can be narrowed down. For example: filter patients with age < 80, or filter patients with data that has been recorded within the last year. The QuantityFilter can be used on a SingleCriterion, e.g. filter all patients that have three records of lab test HbA1c. The AgeFilter uses an additional criterion representing the difference between the date a clinical characteristic has been recorded and the age of the patient in relation to the clinical characteristics.

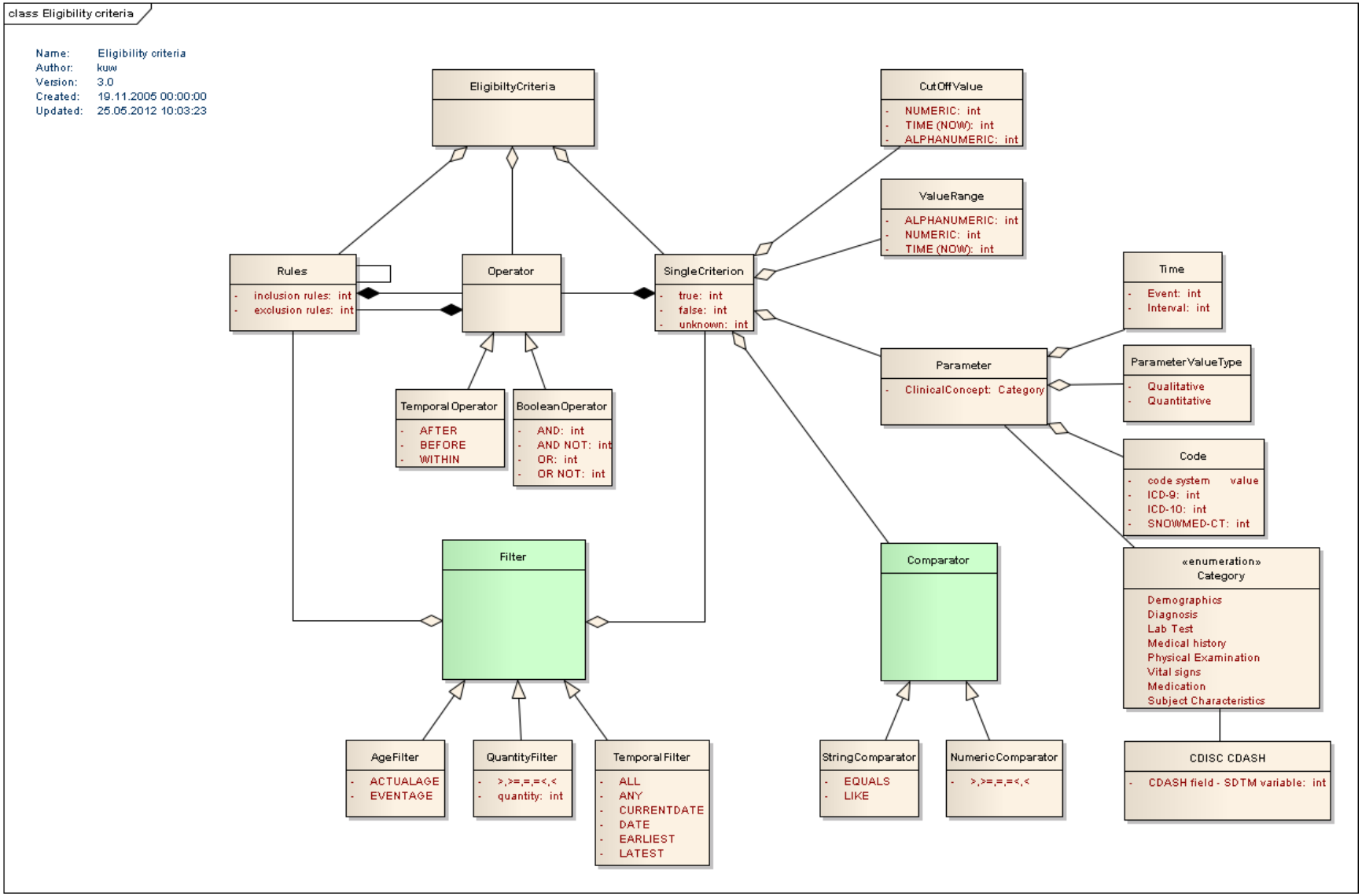


Fig. 29 Eligibility Criteria Information Model (Filter and Comparator are highlighted in green)

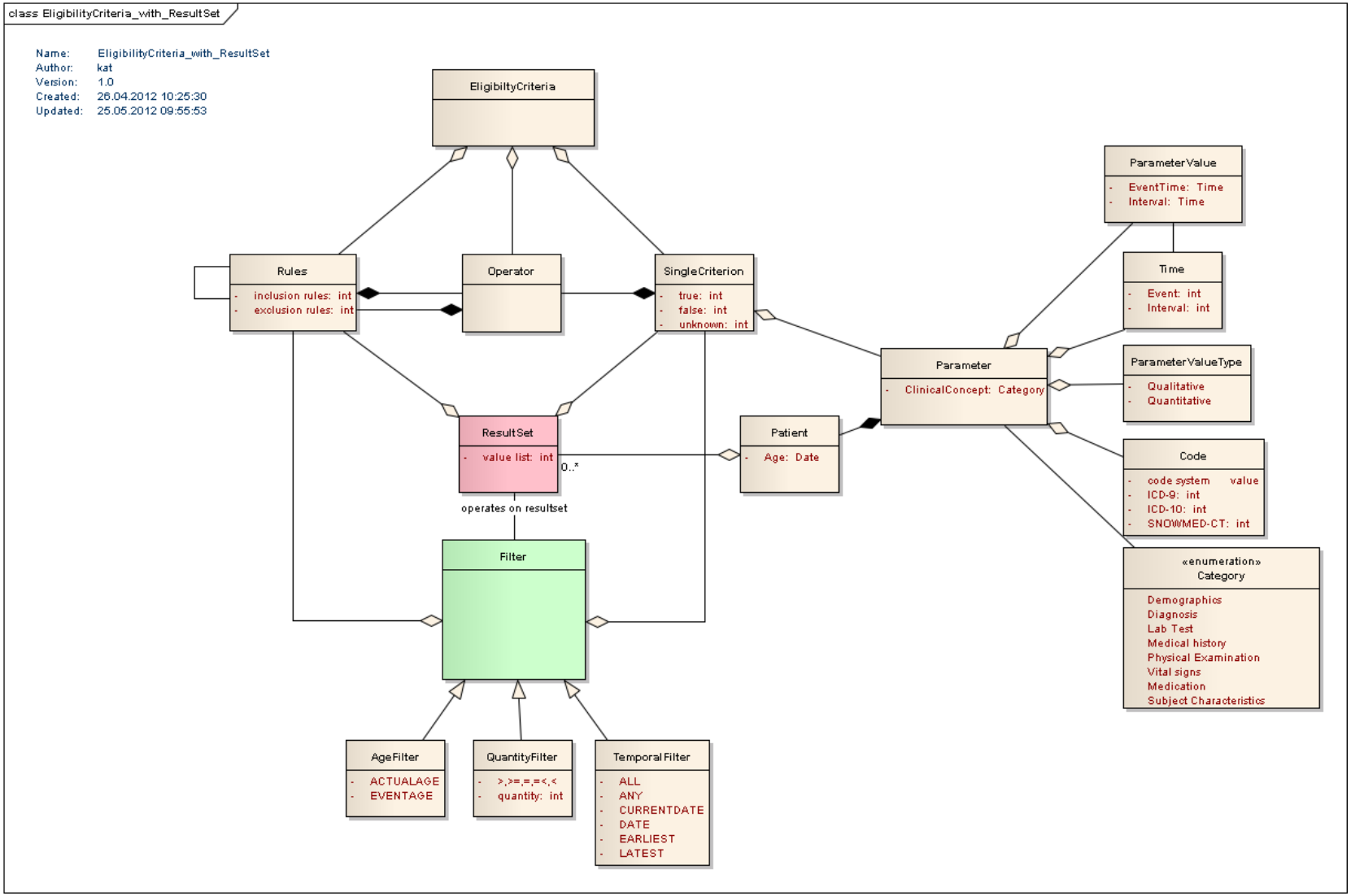


Fig. 30 Eligibility Criteria Information Model with Result Set: the ResultSet (value list) is highlighted in red on which Filters (green) operate

Glossary of eligibility criteria concepts use in ECM

Concept	Sub concept	Description
Code		Classifies medical and health care concepts on the basis of standardised classifications.
Comparator		Defines a “compare function” between two parameters; the result of a comparison should be (true or false).
Comparator	NumericComparator	A function performed on numeric values, such as 1, 2, 3 or 0.5. We propose the following operators to be established: <, ≤, =, ≥, >
Comparator	StringComparator	A function performed on string values, such as diagnosis “equals” “type I diabetes melitus”. We propose the use of “equals”, “like” and “contains” as the basic set of string operators which later may be extended.
CutOffValue		Minimum / maximum values to be taken into account when interpreting values, e.g. lab values of patients: find all patients with HBA1c > 10 %, where 10 is the cut-off value
Eligibility Criteria		The whole set of inclusion and exclusion criteria defined in a specific trial protocol, to be used to query for eligible patients. Eligibility criteria in ECM contain the inclusion and exclusion rules defined in Rules, SingleCriterion and the ResultSet of patients which fulfill the eligibility criteria.
Filter		To screen for parameters/clinical characteristics can be used. We propose to use this filter concept on a more general level.
Filter	QuantityFilter	Can be processed on a SingleCriterion to get a specific quantity value, e.g. all patients who have at least 3 records of lab values HbA1c > 10%.
Filter	AgeFilter	Provides specification of a criterion representing the difference between the date of a clinical characteristic was recorded and the age of the patient in relation to this clinical

		characteristics. For example: find all patients who have been diagnosed with type I diabetes prior the age of 18. Here “EVENTAGE < t” for the SingleCriterion type I diabetes.
Filter	TemporalFilter	The use of temporal filter is important to index events to which other events can be related on a time scale. These filters should cover the identification of the MOST RECENT (FIRST), LAST (LATEST/EARLIEST), ANY or ALL events as index elements.
Operator		Generic concept of Boolean or temporal operators.
Operator	TemporalOperator	SingleCriteria are combined by temporal operators. The following operators are needed as a minimum: AFTER t; WITHIN t and BEFORE t, where t is a value to define a time span (years, months, days, hours or seconds).
Operator	BooleanOperator	There should be no substantial restrictions on the definition of Boolean combinations within Rules, any combination of “AND”, ”OR”, “AND NOT” and “OR NOT” should be possible. All individual inclusion criteria will be automatically linked with “and” and all exclusion criteria with “or.
Parameter		A special kind of variable (e.g. HbA1c) used as a reference to one or more values (ParameterValue).
ParameterValue		Component of the ResultSet of each patient and undergoes further post-processing.
ParameterValueType		Specific type of Parameter, a qualitative (e.g. type I diabetes) or quantitative (e.g. 3) value.
Patient		Part of the ResultSet of an eligibility criteria query with the whole set of associated parameters that has been queried.
ResultSet		Contains the result of a query associated with the EligiblityCriteria. Each result (value list) contains a set of eligible patients and a set of one or more parameters with associated values and timestamps or time intervals.

Rules		Rules are either compositions of logically combined SingleCriterions / Rules or temporally combined SingleCriterions. The usage is restricted to one type of operator (temporal or logical (Boolean) in a single Rules group.
SingleCriterion		Contains the clinical concepts as Parameters and their associated values when a query has been performed. SingleCriterion can be used to compose Rules employing logical and temporal relations.
Time		Specifies either a time interval or a time stamp.
ValueRange		A specified range associated with one parameter to find e.g. all patients with 18 < age of patients < 60.

ECM validation

Examples of inclusion and exclusion criteria were used to validate the ECM. To understand the ECM example criteria were applied to the model; Comparator and Filter were employed to screen for patient numbers. Following two examples for the use of temporal logic and filters were taken from [32].

EXAMPLE 1

How can the group of patients with type I diabetes that have had at least two Glycosylated Haemoglobin (HbA1c) values greater than 10 percent prior the first diagnosis of type I diabetes? Such a query would benefit health care providers who may be managing a group of patients with diabetes, since it readily identifies those patients who have histories of poor glycemic control. The set of all patients who each have had only one HbA1c value greater than 10 percent is less helpful, since many newly diagnosed diabetics will have had such a value at or after diagnosis. Therefore, the exclusion of the first such value would yield a more useful and meaningful group of patients [32].

SingleCriterion

EC-No.	SingleCriterion	Filter	
1	Diagnosis: Type I diabetes mellitus	TemporalFilter	EARLIEST
2	Lab Value: HbA1c > 10%	QuantityFilter	>2

Rules (temporal or boolean composition of SingleCriterion/Rules)

EC-No.	Composition	Operator	
3	EC-No.**2** "BEFORE" EC-No.**1**	TemporalOperator	BEFORE

Results

No.	ResultSet
1	Patients with type I diabetes, only time stamps of the earliest diagnosis of diabetes will be considered (filtered) (e.g. 1224 patients with diagnosis type I diabetes and their associated event times)
2	Only all patients with more than 2 lab values (filtered): HbA1c > 10% are considered (368 patients with at least two lab values (HbA1c > 10%) and their associated event times)
3	Patients from result set No. 1 and No.2 that fulfil the EC-No.3 are in final result set (e.g. 11 patients are considered, when their associated set of event times of EC-No. 1 and EC-No. 2 fulfil the "BEFORE" temporal operation.)

EXAMPLE 2

Find those patients with congenital hypothyroidism which were not seen in the endocrine clinic after their most recent elevated Thyroid Stimulating Hormone (TSH) laboratory result (i.e., a value greater than or equal to 10 µU/ml), the following hypothetic eligibility criteria could be built [32].

SingleCriterion

EC-No.	SingleCriterion	Filter	
1	Diagnosis: congenital hypothyroidism		
2	Lab Value: TSH ≥ 10 µU/ml	TemporalFilter	LATEST
3	Visit	TemporalFilter	LATEST

Rules (temporal or boolean composition of SingleCriterion/Rules)

EC-No.	Composition	Operator
4	EC-No.1 "BEFORE" EC-No.2	TemporalOpearotr
5	EC-No.3 "BEFORE" EC-No.2	TemporalOpearotr
6	EC-No.**4** "AND" EC-No.5	BooleanOpearotr

Results

No.	ResultSet
1	All patients diagnosis equals "congenital hypothyroidism"
2	All patients with latest lab values: TSH ≥ 10 µU/ml
3	All patients from ResultSet 1 where the patients are only considered if the set of all diagnosis timestamps are before the latest lab value timestamp of TSH
4	All patients from ResultSet 3 are considered if the latest visit has been prior to their last recorded TSH ≥ 10 µU/ml
5	All patients that are represented in both ResultSets 4 and 5

12. Discussion and consequences

Clinical trial and research modeling in UML provides a solution to the problem of defining the workflow of clinical trial and research processes, the complex research organizational concepts, care database inclusion and eSource based data collection methods. Because the underlying workflow as required by the TRANSFoRm use cases turned out to be very complex, it was our aim to develop an information model as simple and applicable as possible. Although we based the development of the information model on the highly complex processes of a GCP compliant RCT trial and a case-control study, we were able to extend PCROM with only 17 new objects or items (12 information objects, 3 episode of care related units and 2 areas) to meet all requirements for research in TRANSFoRm. This was achieved by introducing a care area and an ENTRY area into the information model. We choose to enter these two high-ranking concepts to differentiate and at the same time bundle the information requirements of the new objects. Because some of the research operations take place in an overlapping area of care and non-care activities, the new area of "care-related research activities" was introduced. It was gradually extended by the insertion of the concepts for encounter, reason for encounter and episode of care into the area; thus connecting care related research activities with the ICPC standard. The use of a separate area has the advantage that the division concerning the ethics and semantics between patient care and clinical research is highlighted.
Because the information model includes information requirements for a full GCP compliant RCT, for data collection a similar ENTRY area was defined. In the eCRF (ENTRY) area GCP compliant data collection is addressed that connects not only observation, but also assessment activities, intervention activities, and patient reporting with data entry and the associated data semantics. Compared with other information models, TRANSFoRm's CRIM is the only one that covers clinical research in such a comprehensive breadth and at the same time integrates care related research activities like patient appointment, screening and recruitment and is still clearly arranged.
Because the information model should also deliver information requirements for the development of applications, the eligibility concept of CRIM was expanded to an integrated Eligibility Criteria Model (ECM) that specifies the information constraints on the formulation of inclusion / exclusion criteria essential for the development of queries for the Query Tool.

CRIM will be used for the development of TRANSFoRm applications. CRIM specifies the semantic content of information objects involved in clinical research and informs some aspects of the data integration in TRANSFoRM. It is the CDIM [33] which is direct contact with the databases and specifies the data models involved. From its high-ranking position in development of tools and processes and because of its close contact to the clinical research workflow, CRIM can inform the Reference Terminologies and also the security and privacy framework. CRIM will be used in

conjunction with CDIM [33] for the application development to create information artefacts within the TRANSFoRm workbench (fig. 31) including the Query tool and the eCRF / Web questionnaire tools. In addition, it informs the development of study protocols for the TRANSFoRm use cases.

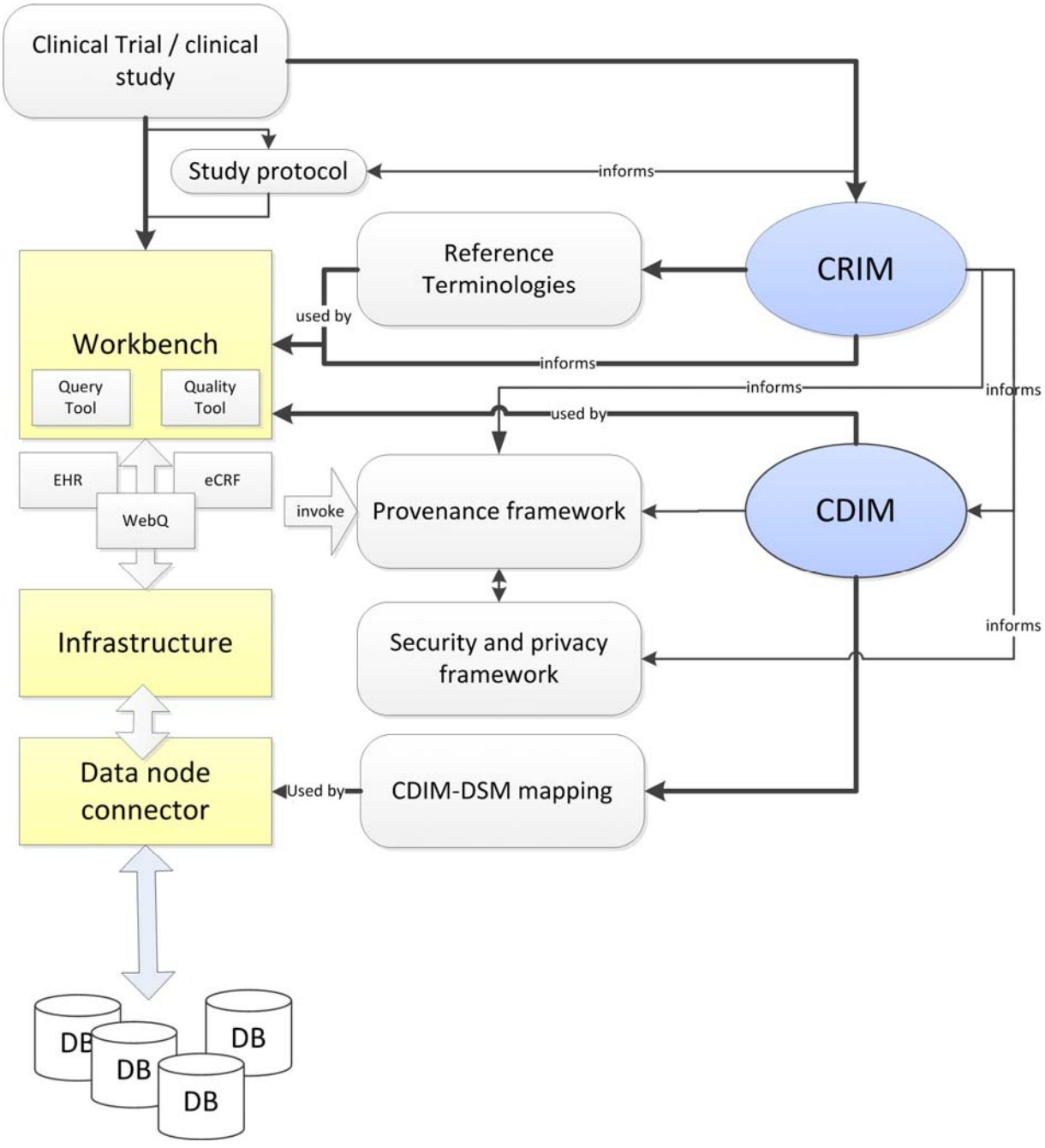


Fig. 31 Dependencies between infrastructure components and models

13. Abbreviations

AE	adverse event
API	application programming interface
BRIDG	Biomedical Research Integrated Domain Group
caBIG	cancer Biomedical Informatics Grid
CDA	Clinical Document Architecture
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CDIM	Clinical Data Integration Model
CRIM	Clinical Research Information Model
CRF	Case Report Form
CRO	clinical research organisation
CTODS	Clinical Trials Object Data System
CTOM	Clinical Trials Object Model
DAM	domain analysis model
dataSHIELD	Data aggregation through anonymous summary-statistics from harmonized individual-level databases
DT2	Diabetes type 2
eCRF	electronic case report form
EC	eligibility criterion
ECG	Electrocardiography
ECM	Eligibility Criteria Model
ECRIN	European Clinical Infrastructures Network
EDC	Electronic Data Capture
EHR	Electronic Health Record
ePCRN	electronic Primary Care Research. Network
ePRO	electronic Patient Reported Outcome
EQ-5D	EuroQol: health-related quality of life
eSDI	eSource Data Interchange
eSource	electronic source data
FDA	Federal Drug Agency
GCP	Good Clinical Practice
GCP IWG	Good Clinical Practice Inspectors Working Group
GO-DARTS	Genetics of Diabetes Audit and Research Tayside
GORD	Gastro Oesophageal Reflux Disease
GP	general practitioner
GPRD	General Practice Research Database

H2-blockers	histamine H2 receptor blockers
HbA1c	glycosylated haemoglobin A1c
HCIT	Health Care Intranet Technologies
HL7	health level 7
HIC	Health Informatics Centre (part of the TAyside medical Science Centre, TASC)
ICH	International Conference on Harmonisation
ICPC	International Classification of Primary Care
NHS	National Health Service
NIVEL	Netherlands institute for health services research
openEHR	open standard specification health data in electronic health records
PHC	Primary Health care Centre
PPI	Proton Pump Inhibitors
PRO	Patient Reported Outcome
PROM	Patient Related Outcome Measures
QoL	Quality of Life
RCT	randomised clinical trial
RIM	Reference Information Model
SDM	Study Design Model
SDTM	Study Data Tabulation Model
SNP	small nuclear polymorphism
TRANSFoRm	Translational Medicine and Patient Safety in Europe
TSH	thyroid stimulating hormone
TTP	Trusted Third Party
UDUS	University of Duesseldorf
UML	Unified Modeling Language
WebQ	web questionnaire
XML	Extensible Markup Language

14. References

[1] Universiteit Antwerpen and Karolinska Institutet: TRANSFoRm WP1/WT1.1: Development of use cases. Deliverable 1, version 2, last updated 2010-09-13

[2] Rosenberg D. and Stephens M.: Use Case Driven Object Modeling with UML. APress, Berkeley, USA, 2007

[3] Speedie SM., Taweel A., Sim I., Arvanitis TN., Delaney B., Peterson KA.: The Primary Care Research Object Model (PCROM): a computable information model for practice-based primary care research. J Am Med Inform Assoc. 2008 Sep-Oct; 15 (5):661-70

[4] http://www.cdisc.org/esdi-document (accessed May 2012)

[5] Payne PRO., Eneida A., Mendonca, M.D. and Starren JB: Modeling Participant-Related Clinical Research Events Using Conceptual Knowledge Acquisition Techniques. AMIA Annu Symp Proc. 2007; 2007: 593–597.

[6] de Carvalho EC., Jayanti MK., Batilana AP., et.al: Standardizing clinical trials workflow representation in UML for international site comparison. PLoS One. 2010 Nov 9; 5(11):e13893

[7] http://www.bpmn.org/

[8] Tina Lee Y. (1999). "Information Modeling from Design to Implementation". Online: http://www.mel.nist.gov/msidlibrary/doc/tina99im.pdf (accessed April 2012)

[9] ICH Topic E 6 (R1): Guideline for Good Clinical Practice. CPMP/ICH/135/95. EMA, London, July 2002

[10] Beale T.: On the possibility of „one information model“ in e-health. OpenEHR: openehr-technical blog (3 May 05, 2011)

[11] Food and Drug Administration: Guidance for Industry. Electronic Source Documentation in Clinical Investigations. Draft. December 2010

[12] GCP Inspectors Working Group (GCP IWG): Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. EMA/INS/GCP/454280/2010. 09 June 2010.

[13] CDISC eSDI Group. Leveraging the CDISC standards to facilitate the use of electronic source data within clinical trials, CDISC; 2006.

[14] Ohmann C., Kuchinke W. (UDUS), van Veen E.-B. (MedLawconsult), Verheij R. (NIVEL), Farrell S. (TCD): TRANSFoRm. Report on regulatory requirements, confidentiality and data privacy issues. Part B: Confidentiality and data privacy framework (Deliverable 3.2). WP3, Deliverable 3.2. Version 1 (31 March 2011)

[15] Bell A.: Data Anonymisation and Linkage. Health Informatics Centre (HIC). University of Dundee. PowerPoint Presentation - National e-Science Centre. www.nesc.ac.uk/talks/.../SINAPSE_HIC_Bell.ppt (accessed April 2012)

[16] Wolfson M., Wallace SE., Masca N., Rowe G, Sheehan NA., Ferretti V., LaFlamme P., Tobin MD., Macleod J., Little J., Fortier I., Knoppers BM., Burton PR.: DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual - level data without

sharing the data. Int J Epidemiol. 2010 Oct; 39 (5):1372-82
[17] Fassaert T., Nielen M., Verheij R., Verhoeff A., Dekker J., Beekman A., de Wit M.: Quality of care for anxiety and depression in different ethnic groups by family practitioners in urban areas in the Netherlands. General Hospital Psychiatry: 2010, 32, 368-376
[18] http://www.healthcit.com/HCIT/clinical-trials-object-data-system
[19] BRIDG Model (Release 3.1), Biomedical Research Integrated Domain Group (BRIDG), 29 February 2012.
[20] BRIDG Model. Release 3.1 User's Guide. 29 February 2012. Biomedical Research Integrated Domain Group (BRIDG)
[21] Peterson KA., Fontaine P., Speedie S.: The Electronic Primary Care Research Network (ePCRN): a new era in practice-based research. J Am Board Fam Med 2006; 19:93-97
[22] Clinical Data Interchange Standards Consortium: CDISC Study Design Model in XML (SDM-XML), Release Version 1.0 (2011)
[23] CDISC CDASH Project Team: Clinical Data Acquisition Standards Harmonization (CDASH). CDASH_STD-1.1 (19 April 2010)
[24] Beeler G W.: Introduction to HL7 RIM. Presentation. Online available: http://www.hl7.org/documentcenter/public_temp_8423DD4D-1C23-BA17-0C2F472AE6B28AB5/calendarofevents/himss/2009/presentations/Reference%20Information%20Model_Tue.pdf (accessed April 2012)
[25] The openEHR Reference Model: EHR Information Model. Editors: Beale T., Heard S., Kalra D., Lloyd D. Revision 5.1.0. Date of issue: 08 Apr 2007. The openEHR Foundation (2007)
[26] http://www.en13606.org/the-ceniso-en13606-standard
[27] Schloeffel P., Beale T., Hayworth G., Heard S, Leslie H.: The relationship between CEN 13606, HL7, and openEHR. In: HIC 2006 and HINZ 2006 Proceedings (Internet); Warren J. (Editor). Brunswick, Vic.: Health Informatics Society of Australia, 2006
[28] Beale T.: ISO 13606 2012 revision openEHR proposal. 9 Mar 2012, openEHR.org dashboard. Available online: http://www.openehr.org/wiki/display/stds/ISO+13606+2012+revision+openEHR+proposal (accessed: May 2012)
[29] Austin T., Kalra D., Tapuria A., Lea N., Ingram D.: Implementation of a query interface for a generic record server. Int J Med Inform. 2008 Nov; 77(11):754-64
[30] Ohmann C., Karakoyun T. and Kuchinke W.: Handling of temporal relations in EHR4CR. EHR4CR document draft 1 (30. Sept. 2011)
[31] Ross J., Tu S., Carini S., and Sim I.: Analysis of Eligibility Criteria Complexity in Clinical Trials. AMIA Summits Transl Sci Proc.; 2010: 46 – 50
[32] Nigrin D J., Kohane I S.: Temporal expressiveness in querying a time-stamp-based clinical database. JAMIA. 2000; 7: 152-163
[33] Ethier J-F., Dameron O., McGilchrist M., Burgun A.: TRANSFoRm: Clinical Data Integration Model (CDIM), Version: V1.0, 19.04.2012