ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP
Contract Nr: 288048

Deliverable: D8.4
Specifications of the evaluation and validation scenarios and demonstrators for the clinical pilots

Due date of deliverable: (31-01-2015)
Actual submission date: (03-04-2015)

Start date of Project: 01 February 2012 Duration: 42 months

Responsible WP: FORTH

Revision: <proposed>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Lefteris Koumakis	FORTH	koumakis@ics.forth.gr
Haridimos Kondylakis	FORTH	kondylak@ics.forth.gr
Evangelia Maniadi	FORTH	maniadi@ics.forth.gr
Maria Psaraki	FORTH	psaraki@ics.forth.gr
Manolis Tsiknakis	FORTH	tsiknaki@ics.forth.gr
Alan Dahi	LUH	dahi@iri.uni-hannover.de
Magdalena Góralczyk	LUH	goralczyk@iri.uni-hannover.de
Pascal Coorevits	EuroRec	pascal.coorevits@eurorec.org
Jasper van Leeuwen	Philips	Jasper.van.leeuwen@philips.com
Kristof De Schepper	Custodix	kristof.deschepper@custodix.com
Kerstin Rohm	FhG IBMT	kerstin.rohm@ibmt.fraunhofer.de
Stefan Rueping	FhG IBMT	stefan.rueping@iais.fraunhofer.de
Cyril Krykwinski	IJB	cyril.krykwinski@bordet.be
Monique Hendriks	Philips	monique.hendriks@philips.com
Sheng Yu	UOXF	sheng.yu@oncology.ox.ac.uk
Francesca Buffa	UOXF	francesca.buffa@oncology.ox.ac.uk
Norbert Graf	UdS	Norbert.Graf@uniklinikum-saarland.de
Wytze Vlietstra	VUA	w.j.vlietstra@vu.nl
Jeroen Keijser	Philips	jeroen.keijser@philips.com
Scott Marshall	Maastro	m.scott.marshall@maastro.nl
Keyur Mehta	GBG	Keyur.Mehta@germanbreastgroup.de

0.2 Documents history

Document version #	Date	Change
V0.1	01/11/2014	Starting version, template
V0.2	15/11/2014	Definition of ToC
V0.3	20/02/2015	First complete draft
V0.4	12/03/2015	Integrated version (send to WP members)
V0.5	15/03/2015	Updated version (send PCP)
V0.6	19/03/2014	Updated version (send to project internal reviewers)
Sign off		Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.3 Document data

Keywords	

Editor Address data	Name: Lefteris Koumakis Partner: FORTH Address: N. Plastira 100 Vassilika Vouton Heraklion Phone: +30 2810 391424 Fax: +30 2810 391448 E-mail: koumakis@ics.forth.gr
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

1 INTRODUCTION 5
1.1 Method 6
1.2 Structure of the Deliverable 8
2 PERSONAL MEDICAL INFORMATION RECOMMENDER 9
3 DATA MINING OF CONSULTATION 14
4 CONTEXTUALIZED OVERVIEW 18
5 TRIAL RECRUITMENT 22
6 PROTOCOL FEASIBILITY 26
7 UPDATE OF GUIDELINES 28
8 HYPOTHESIS GENERATION 31
9 OUTCOME PREDICTION 34
10 DIAGNOSTIC CLASSIFIER 37
11 PREDICTION OF SAES/SUSARS 40
12 PATIENT DIARY & LONG-TERM FOLLOW-UP 45
13 AUTOMATIC DETECTION AND REPORTING OF SAES/SUSARS 49
14 MICROBIOLOGY SAE 60
15 REPORTING EPISODES OF FEBRILE NEUTROPENIA 63
16 CANCER REGISTRY AND TUMOUR BANK REPORTING 66
17 CONCLUSIONS 69
18 REFERENCES 75
APPENDIX A 76

1 Introduction

For EURECA, validation & evaluation does not refer only to software components but also to processes such as clinical scenarios. The validation covers all aspects of the process including the EURECA environment, any hardware that the environment uses, interfaces to other systems, the users, training and documentation as well as the management of the system and the validation itself after the system is put into use.

We intend to perform an extensive evaluation of the clinical services offered by the infrastructure. The idea is that measurable parameters will be established in cooperation with the responsible clinical partners (e.g. recruitment rate, the number of SAE/SUSAR avoided etc.) for each clinical service offered within EURECA. Those measurable parameters could be monitored for a time frame [x1, x2] where the EURECA infrastructure is not used. Then, EURECA services could be used and the same parameter would be monitored. In this way the real impact of the EURECA infrastructure could be demonstrated. A simplified example of such a procedure is shown in the following figure. However, due to lack of time this comparative, evaluation approach cannot be followed.
The evaluation will be done by measuring the established parameters of each clinical scenario.

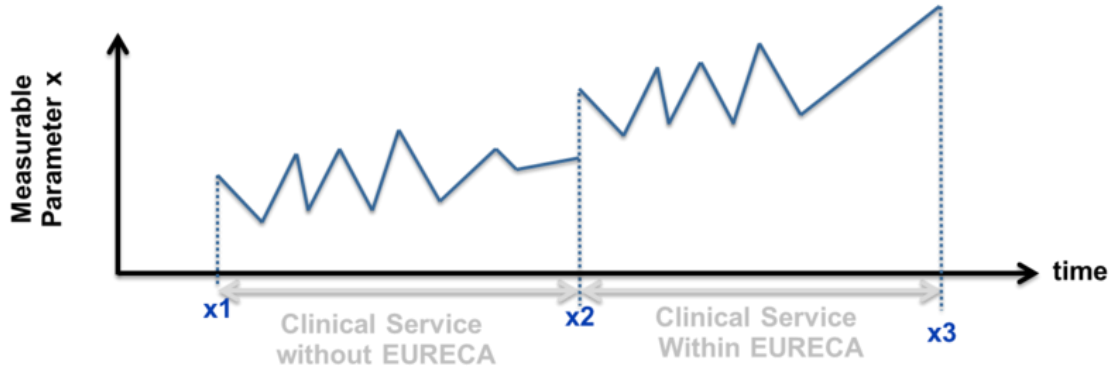


Figure 1: Evaluation procedure for the clinical services offered by EURECA.

End-user evaluation of the EURECA infrastructure will be conducted through 15 selected scenarios covering the anticipated usage of the infrastructure, from administration of the software components to specific clinical tests. The end users who will participate at the evaluation phase will also fill in an evaluation form for each EURECA component. The evaluation form will cover all the appropriate quality characteristics from the product quality model of the ISO/IEC 25000 series.

At the evaluation phase different type of users, such as physicians, system developers and patients will participate. Having such a diverse target group of evaluators, the evaluation forms must be:

- simple
- accurate
- easy to understand (especially for non IT experts)
- non time consuming
- without loss of functionality/quality

For that reason we translated the crucial sub-characteristics of software quality measures into simple questions (in natural language). The evaluation form of EURECA is a list of such questions where the evaluator will answer with a degree of satisfaction with Likert scale. Likert scale is based on forced-choice questions, where a statement is made and the respondent then indicates the degree of agreement or disagreement with the statement on a 5 point scale. The Generic Evaluation Questionnaire consists of two forms:

- The selected sub-characteristics, for the evaluation form of the EURECA scenarios and components, and its translation into a simple question for the end user can be found in the Generic Evaluation Questionnaire Form A in deliverable D8.1.
- We also use the System Usability Scale (SUS) for global assessment of systems usability. The SUS can be found at the Generic Evaluation Questionnaire Form B in deliverable D8.1.

The evaluation process in EURECA takes the end-users (clinicians, bioinformaticians, research nurses, data managers, epidemiologists) into its focus for testing the developed tools and the integrated clinical scenarios.

1.1 Method

This deliverable reports on the evaluation and validation procedures for the EURECA scenarios. Specifically, we first identify and report the status of the 15 clinical scenarios identified in the requirement elicitation phase of the project [1] [2], shown also in Table 1. This document describes the evaluation procedure to be followed for each one of these clinical scenarios.

Table 1: The list of the clinical scenarios

Evaluation Leader	Scenario	Development leader
UdS	Personal medical information recommender	FORTH, FhG IAIS
	Data mining of consultation	FhG IAIS
	Prediction of SAEs/SUSARs	Philips
	Automatic detection and reporting of SAEs/SUSARs	FhG IBMT
	Microbiology SAE	FhG IBMT
UOXF	Contextualized overview	VUA
	Patient Diary & Long-term follow-up	FORTH
	Hypothesis generation	UOXF
	Outcome prediction	UOXF
	Diagnostic classifier	UOXF
IJB	Reporting episodes of febrile neutropenia	IJB
	Cancer registry and tumour bank reporting	IJB

Maastro	Update of guidelines	VUA
	Trial recruitment	Philips
GBG	Protocol feasibility	Custodix

For each one of these scenarios, the objective of this deliverable is to establish the procedures for answering the following question "*Does the software do something of sufficient value?"* i.e. to demonstrate that the system has a positive benefit to its users.

To be able to answer such a question a number of parameters should be specified first. To help in identifying and elaborating on these parameters a form has been created to be filled for each clinical scenario, shown in Appendix A. The form foresees several questions that should be answered for each scenario. These questions are grouped in the following categories.

- **General Information**: In this category the technical and the clinical evaluation leader should be specified and a detailed description of each scenario should be provided.
- **Setup Information**: Then the location of the evaluation should be specified, and the setup details. For example who will be the evaluators, how many of them will be used, what will be the duration and the time plan of the evaluation.
- **Data Information**: Apart from these, the data that will be used for the evaluation of each scenario should be clearly specified. To get access to data the CDP servers might need to be accessed and/or other prospective data as well.
- **Evaluation Type**: Then a key question is whether the evaluation will be a "*Proof of concept"* or a full-fledged clinical evaluation.
    - **Proof of concept**: In case of proof of concept evaluation, the specific parameter that will be evaluated should be defined and the method for evaluating it as well.
    - **Clinical Evaluation**: In case of Clinical Evaluation, firstly the type should be specified (sequential, parallel or retrospective) and then the inclusion and exclusion criteria and the number of patients/clinicians that will be used. The evaluation hypothesis and the control group should be clearly defined and the appropriate metrics as well. Every quantifiable feature of software and every quantifiable interaction of software with its environment that correlates with a characteristic can be established as a metric. Metrics can differ depending on the environment and the different end user groups. The task will be to produce values that are (1) formal enough to serve as a basis for comparison amongst alternative methods under consideration; (2) mappable to utility e.g., measuring the weight of some object of evaluation should only happen, if it is clear how weight relates to utility. Finally there should be clear assessment criteria to contribute to the assessment of the quality of each scenario.

The results of evaluation of the aforementioned scenarios will be presented at D8.5 Report on the evaluation and validation of the EURECA environment and services and finally at D8.5 there will be also a report on the user workshops at the different clinical sites.

1.2 Structure of the Deliverable

The remaining of this document is structured as follows: Chapter 2 to 16 reports the evaluation procedures for each one of the aforementioned clinical scenarios. Then in Chapter 17, we summarize the results of this deliverable and provide directions for our next steps.

2 Personal medical information recommender

The Personal Medical Information Recommender (PMIR) is a tool that allows clinicians and patients to obtain objective information (e.g. treatments) about the patient/disease. The Personal Medical Information Recommender is a web-based tool. UdS as the evaluator at the Hospital for Paediatric Oncology and Haematology has access to this tool via a secure website.

The Personal medical information recommender is developed by FORTH using Indivohealth™. A patient or the treating physician wants to find relevant information or literature about the medical situation of him/her or of one of his/her patients. Actual data from the patient are presented in the patients' EHR (Indivohealth™), which is shown in Figure 2. John Smith is the fictive patient with his medical data stored in Indivohealth™ used in the evaluation of the PMIR.

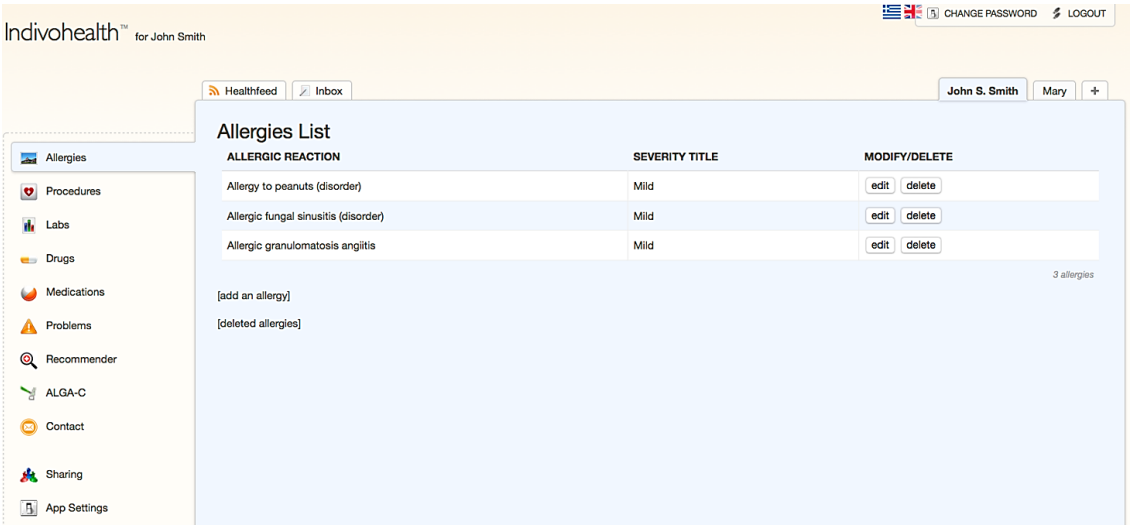


Figure 2: EHR (Indivohealth™) of John Smith showing his allergies.

On the left side of the EHR (Indivohealth™) the different tools that the patient can use are displayed, presenting specific health data, like allergies, procedures, patient's medication or lab data, etc. All kinds of health problems can be added to his EHR. He can add, edit or delete them. In addition, he can use tools to analyse his data or check interactions between drugs. He can fill in the ALGA questionnaire or send messages to other people, including his physician. One further tool is the PMIR (Personal Medical Information Recommender) that can be used by the patient or the treating physician.

In addition the patient can share his own health data with others in care networks (Figure 3).

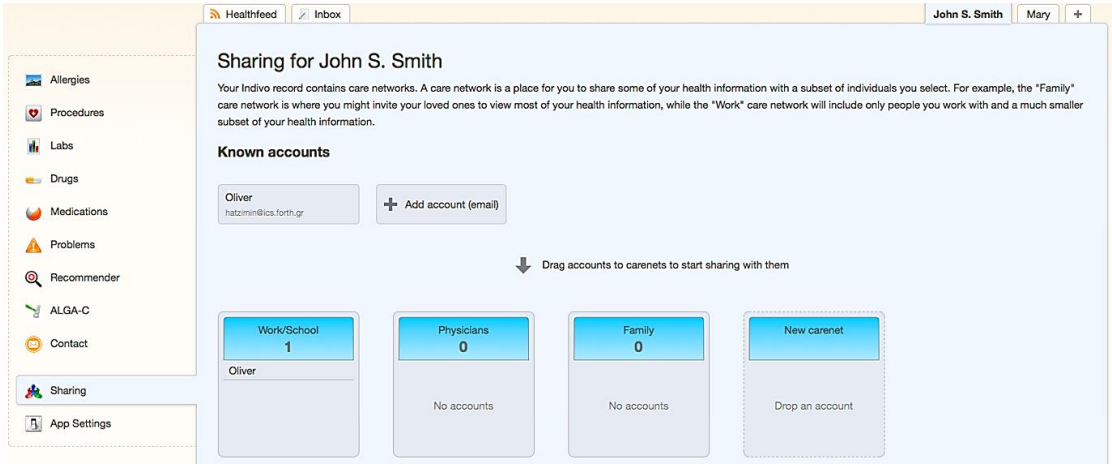


Figure 3: EHR (Indivohealth™) showing the care network that can be easily used by the patient by dragging accounts to different carenets.

The patient can limit his/her carenets' access by choosing which data to share with whom (Figure 4).

Figure 4: EHR (Indivohealth™) showing how to limit carenet's access to specific data.

The PMIR, as one of the tools provided in EHR (Indivohealth™), is shown in Figure 5.

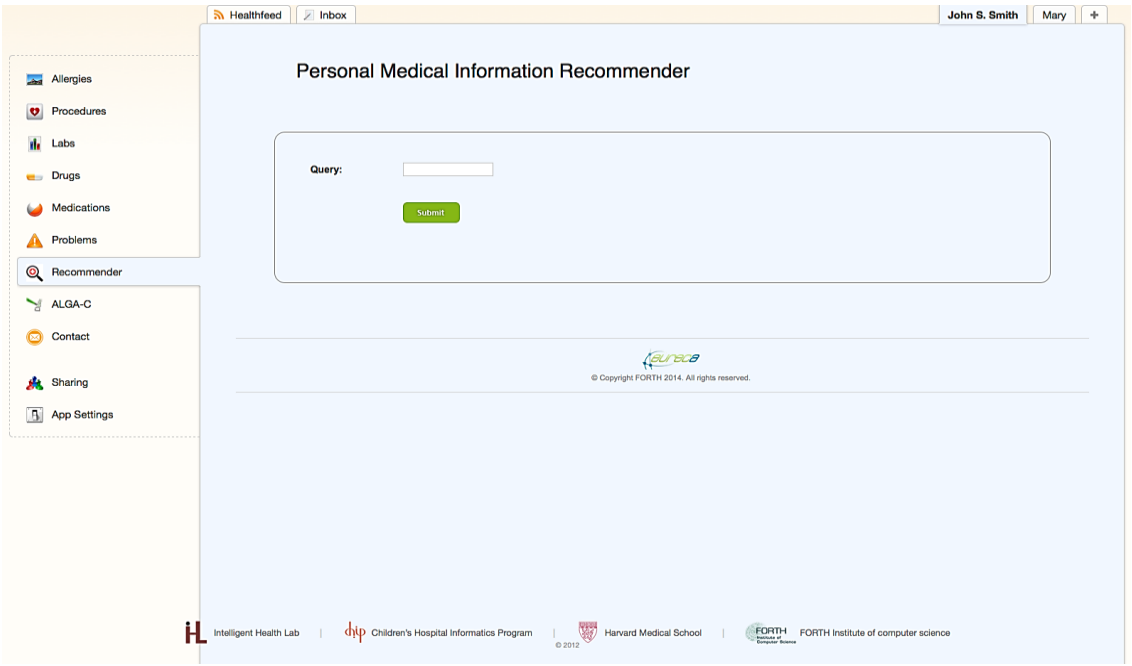


Figure 5: EHR (Indivohealth™) showing the PMIR.

Within the PMIR it is possible to ask a question in natural language, like: "*What is the best treatment for my disease?*" By submitting this question into the PMIR 354 papers are listed, which are ranked according to the relevance for the patient (Figure 6).

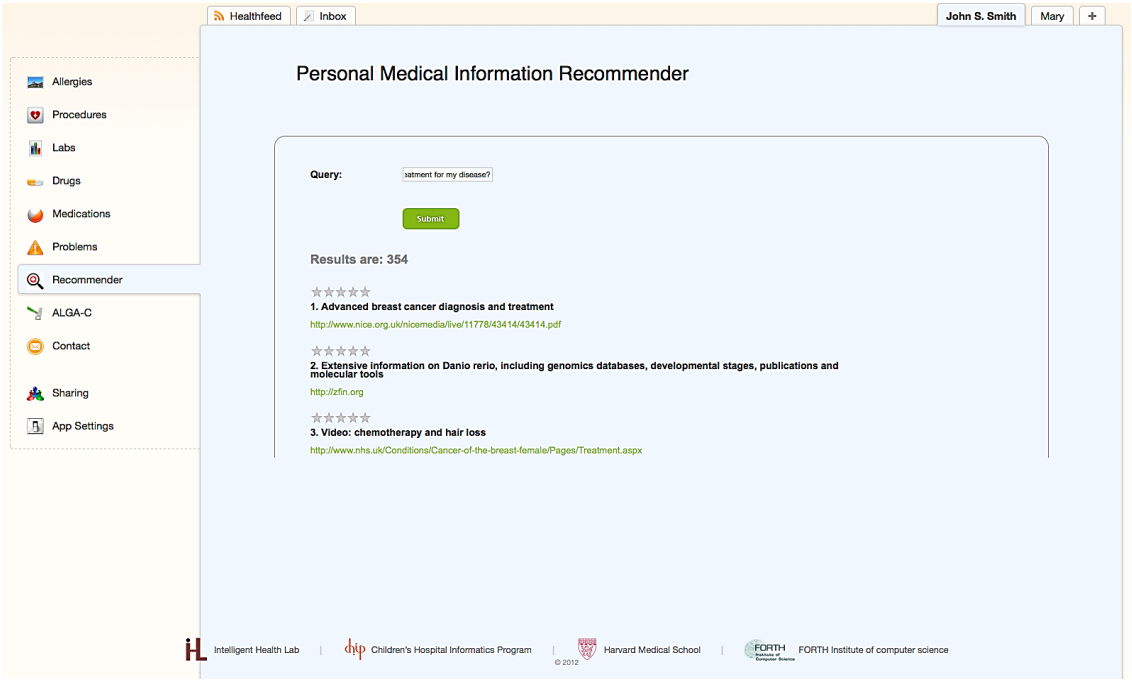


Figure 6: EHR (Indivohealth™) showing the answer in PMIR to the question: "What is the best treatment for my disease?".

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 2: Main features of the evaluation procedure for the "Personal medical information recommender" scenario.

		Description
General	Evaluation Scenario	Personal Medical Information Recommender
	Evaluation Leader	Norbert Graf
	Description of the scenario	The Personal Medical Information Recommender (PMIR) is a tool that allows clinicians and patients to obtain objective information about the patients, the diseases and the treatments.
	Description of the evaluation procedure	The Personal Medical Information Recommender is a web-based tool. UdS as the evaluator at the Hospital for Paediatric Oncology and Haematology has access to this tool.
Setup	Location of evaluation	UdS - Hospital for Paediatric Oncology and Haematology
	Setup details	
	Evaluator's expertise	Professor for Paediatric Oncology Haematology
	Number of evaluators	At least 3 physicians and 3 study nurses
	Duration of the evaluation	This will be done up to the end of March.
	Time plan	Beginning of 2015 after the installation of the system
Data	Kind of data used for the evaluation	Fake data of patients entered in the EHR (Indivohealth™)
	For retrospective data	Only fake data are used
	If you plan to access CDP servers	Not planned
	If you plan on using prospective data	Not planned
	Type(s) of evaluation	The evaluation will be a proof of concept but also asking the question, if the tool can be used in clinical care.
Proof of Concept	What exactly will be evaluated?	It will be evaluated how good the medical information for a given patient will be.
	How it will be evaluated?	The templates given in D8.1. will be used.
	How many clinicians/users will be involved?	At least 3 clinicians and 3 study nurses.

		How many times will they fill-in the usability questionnaires?	If the tool is not satisfying the end-users there will be an iterative process with the developers to optimize the tool and a new evaluation will be done by the same end-users.
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation in an iterative process until the tool is satisfying all end-users. It will take place at the same location (UdS) with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Informed consent is not needed as it will not have any impact on patients. 10 fake patients with different diseases will be evaluated.
		Inclusion/Exclusion Criteria?	No exclusion criteria. There will be 10 fake patients with different diseases.
		Will the EURECA tool be compared with the tool you normally use?	We do not use such a comparable tool.
		Which is the evaluation hypothesis?	The hypothesis is that the tool is helpful in predicting recommendations for patients with different diseases.
	Evaluation factor 1	Quality metrics	Comparing the prediction with the recommendation of a clinician.
		Measured value	The question to be answered will be: “Is the information provided by the tool helpful?” The answer will be binary (YES or NO) with comments to justify the selection.
		Rating levels	The rating given in D8.1 will be used.
	Criteria	Assessment criteria	Assessment criteria are the usability during clinical practice.

3 Data mining of consultation

A local physician asks for consultation by filling in a consultation request form. This form includes data regarding the history of the patient, the treatment given so far, side effects of treatment, the actual situation of the patient and a concrete question. The consultation request is sent by the treating physician via email to the trial chairman of SIOP 2001. If incomplete information is provided the trial chairman will request more details by the local treating physician. After receiving all information needed to give a recommendation, this recommendation is sent by email to the local physician. The request of the local physician and the answer of the trial chairman are stored in the database of the SIOP 2001 trial mainly for documentation purposes, meaning that during the trial SIOP 2001 consultations of patients are recorded and stored in the clinical trial database.

For analysing and using these data from the trial database, these retrospective data are anonymized by Custodix and proven by LUH as being anonymous. The purpose of the data mining scenario is to use these retrospective data to find answers to new requests for consultation. This is achieved by finding similar questions from the existing database, such that (a) frequently asked questions (FAQs) can be identified, and (b) the known answers to historical similar questions can be used as a suggested answer to the new question. The evaluation of the tool will be done in two steps: first, it will be evaluated whether the tool is able to identify FAQs in the historical dat. Second, it will be evaluated whether the assignment of a new question to an already identified FAQ or to the most similar known question is correct from the medical point of view.

Identification of FAQs: the tool will be run on the database of existing questions and produce

(1) a list of topic that frequently appear in the texts. Each topic consists of a list of descriptive keywords. In addition, the tool will give information on the dependency of these topics.
(2) A list of candidates for FAQs

The evaluator will evaluate the results of the tool under the following criteria

(1) For each topic in the list, the reviewer will quantify the understandability of the topic on the range of 1-3, 1 being not understandable, 2 being somewhat understandable, and 3 being clearly understandable. A topic shall be judged to be understandable (2 or 3), if the evaluator can give a meaningful name or 1-sentence description on what which aspect of the consultation this topic refers to. For a meaningful topic, this label plus an assessment of the relevance of this topic to understand the content of the consultation query (range 1-3, 1 not being relevant, 2 somewhat relevant and 3 relevant) shall be recorded.
The overall quality of the topic list shall be measured by the average understandability of the topics, the number of understandable and the number of understandable or somewhat understandable topics, and the number of relevant or somewhat relevant topics amongst all understandable topics.
(2) For each candidate FAQ, the evaluator shall judge whether this consultation query is indeed a frequently asked question, in the sense of whether he receives this or a similar question at least once a month (frequently asked) or

at least once a year (somewhat frequently asked). The evaluator will also evaluate whether this candidate FAQ too specific to be an FAQ, too general to be an FAQ, or neither too general nor too specific.
The overall quality of the candidate FAQ list shall be measured by the number of frequent or somewhat frequent questions in the list, plus the number of too general or too specific candidates.

If the quality of the evaluation is not sufficient, an iterated optimization will be performed.

Evaluation of the assignment of historical questions to new queries: a new query will be input to the tool. The tool will produce

(1) A relevant FAQ, if available, or a message that not relevant FAQs could be found
(2) The most similar existing query from the database.

The evaluator will assess whether the FAQ is relevant, and whether the existing query is indeed similar. The test can be executed using a new query, or by splitting up the data set into a training and a test set. The overall quality of the assignment of historical questions to new queries shall be measured by (1) the fraction of recommended FAQs that are indeed relevant, and (2) the number of recommended similar queries that are indeed similar.
This evaluation will only be performed once the identification of FAQs is sufficiently solved.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 3: Main features of the evaluation procedure for the "Data mining of consultation" scenario.

		Description
General	Evaluation Scenario	Data Mining of consultation
	Evaluation Leader	Norbert Graf
	Description of the scenario	A local physician asks for consultation by filling in a consultation request form.
	Description of the evaluation procedure	During the trial SIOP 2001 consultations of patients are recorded and stored in the clinical trial database. These retrospective data are anonymized by Custodix and proven by LUH as being anonymous. These retrospective data will be used to (1) identify FAQs, (2) identify similar historical queries to new queries. The results will be checked by the trial chairman of SIOP 2001, rating them as helpful or not.
Setup	Location of evaluation	UdS - Hospital for Paediatric Oncology and Haematology
	Setup details	The data mining tool needs to be installed at UdS or a remote access will be established to

		IAIS.
	Evaluator's expertise	Professor for Paediatric Oncology Haematology
	Number of evaluators	The trial chairman of SIOP 2001
	Duration of the evaluation	This will be done within 4 weeks after setup of the system.
	Time plan	Beginning of 2015 after the installation of the system.
Data	Kind of data used for the evaluation	Retrospective and anonymized data of the consultation CRF of SIOP 2001
	For retrospective data	The evaluation will be done with anonymous data.
	If you plan to access CDP servers	Not planned
	If you plan on using prospective data	Not planned
	Type(s) of evaluation	The evaluation will be a proof of concept but also asking the question, if the tool can be used in clinical care.
Proof of Concept	What exactly will be evaluated?	It will be evaluated how good the data mining tool will identify FAQs and relevant similar queries using the quality measures identified abouce.
	How it will be evaluated?	The templates given in D8.1. will be used.
	How many clinicians/users will be involved?	Only the trial chairman
	How many times will they fill-in the usability questionnaires?	If the tool is not satisfying there will be an iterative process with the developers to optimize the tool and a new evaluation will be done by the same end-users using the same questions.
Clinical Evaluation **Set-Up**	Clinical Evaluation Type	Same end-user will do the evaluation in an iterative process as long as the tool is not satisfying. It will take place at the same location with the same parameters.
	How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Informed consent is not needed. Only retrospective and anonymized data will be used.
	Inclusion/Exclusion Criteria?	No exclusion criteria.
	Will the EURECA tool be compared with the tool you normally use?	We do not use such a comparable tool. To our knowledge there are no comparable tools available.
	Which is the evaluation	The hypothesis is, that the tool is helpful in

		hypothesis?	correctly identifying FAQs and relevant historical queries.
	Evaluation factor 1	Quality metrics	(1) Average understandability of topics, the number of understandable and the number of understandable or somewhat understandable topics, and the number of relevant or somewhat relevant topics amongst all understandable topics. (2) Number of frequent or somewhat frequent questions in the list, plus the number of too general or too specific candidates. (3) The fraction of recommended FAQs that are indeed relevant, and the number of recommended similar queries that are indeed similar.
		Measured value	Rating 1-3 with the levels of (no, somewhat, yes) for all questions
		Rating levels	Rating will be done according to the template given in D8.1.
	Criteria	Assessment criteria	Assessment criteria are the usability during clinical practice.

4 Contextualized overview

The physician facing contextualization task aims to aid physicians in their information retrieval tasks during a patient – physician consultation. By integrating the patient data contained in the EHR, results of queries can be specialized to the patient's specific context, which is thought to enhance efficiency and quality of care.

The evaluation will start with a demonstration that aims to provide a guidance for the evaluators and followed by a number of tasks to complete. The following steps detail evaluation plan of the EURECA tool.

1. Discuss the possible evaluation dates and provide background information for potential evaluators.
2. Install the Contextualisation tool on the evaluation site, which needs to be decided with the evaluators.
3. In the first part of the evaluation, the evaluators will be given a demo about the system. Particularly the staff will show the users how to get a contextualised overview of the patient.
4. In the second part, the evaluators will perform the same task, i.e browsing patient information, and evaluate whether the retrieved documents are relevant or not.
5. The observation of the activity of the evaluators will be recorded to perform a Keystroke-level model (KLM) analysis.
6. Goals and sub tasks will be established to be performed, e.g. ask the clinician to "start with one patient to find out relevant information".
7. The mouse click will be recorded as a BB ('button press-button release') event. The average clicks will be recorded to compare with free text search engine such as Google.
8. After the exercise, the evaluators will answer the evaluation questionnaire.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 4: Main features of the evaluation procedure for the "Contextualized overview" scenario.

		Description
General	Evaluation Scenario	Contextualised overview
	Evaluation Leader	UOXF
	Description of the scenario	The physician facing contextualization task aims to aid physicians in their information retrieval tasks during a patient – physician consultation. By integrating the patient data contained in the EHR, results of queries can be specialized to the patient's specific context, which is thought to enhance efficiency and quality of care.
	Description of the evaluation procedure	Evaluation will be held inside the NHS UK. VUA will provide with the contextualisation

			application. Real patient data will be imported to the system (via the EURECA service and database import). Clinicians from NHS will use & evaluate the tool.
Setup	Location of evaluation		Oncology department of Oxford University
	Setup details		The tool will be installed on a VU server and publically available, although an EURECA SIL login is required. The patient data will be hosted by UPM.
	Evaluator's expertise		A clinician.
	Number of evaluators		3 or 4
	Duration of the evaluation		Half an hour
	Time plan		Mid. – end of April 2015
Data	Kind of data used for the evaluation		Retrospective
	For retrospective data		Access the CDP servers to perform the validation (the server of UPM, or of Custodix).
	If you plan to access CDP servers		(a)Yes (b)Yes
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		The functionality and the usability of the contextualisation application will be evaluated by the clinicians.
	How it will be evaluated?		The clinicians will have 30 minutes to evaluate the tool. In the first 10 minutes the evaluators will see a short demonstration about how the contextualisation tool works. Then in the other 10 minutes they will be guided to use the tool, e.g., viewing patient information and the relevant information retrieved by the tool based on the context of the patient data. After that, they will answer the questions on the questionnaire and comment on the system in the remaining time.
	How many clinicians/users will be involved?		Three
	How many times will they fill-in the usability questionnaires?		Once
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation. It will take place at the same location with the same parameters.

		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Each clinician will test around 1-3 patients. The patient data will be released, and therefore will not require informed consent.
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool compared with the tool you normally use?	None
		Which is the evaluation hypothesis?	With the EURECA contextualisation tool it is possible to provide relevant documents and information based on patient data
	Evaluation factor 1	Quality metrics	For filtering use cases, the application is either incapable of returning a result, or it returns a correct result. As such, the number of cases in which it the application is capable of returning an answer is measured.
		Measured value	Recall
		Rating levels	
	Evaluation factor 2	Quality metrics	For the ranking use cases, a comparison in time will be made between the manual formulation of a query, and the querying process the system will execute. Qualitative assessment will be described below
		Measured value	Time
		Rating levels	
	Evaluation factor 3	Quality metrics	For the ranking use cases, there is no golden standard to compare the generated results to. Therefore a more subjective comparison will be made by measuring the number of titles on which a user clicks up to a certain number, after both a manual formulation of a query and the query process executed by the system
		Measured value	The expert examines the first n (n = 15?) titles, and clicks the hyperlinks opening the articles (s)he considers potentially relevant. This operation is both executed in the contextualization application and in PubMed, in which he/she formulated their own query. The subjective relevance of results between these two processes are compared.
		Rating levels	

	Criteria	Assessment criteria	Usability assessment based on ISO/IEC 25000 series and on SUS scale

5 Trial Recruitment

The trial recruitment clinical scenario is the basis for the Yakobo clinical trial recruitment tool. Yakobo uses a set of web services provided by the EURECA platform and with a visual and touch based UI that shows which patients are potential candidates for a clinical trial. The purpose of the validation of Yakobo is to show it to people who deal with trial recruitment in the clinical domain and see whether this visual approach of Yakobo's design is an understandable and effective manner to support clinical trial recruitment.

The MAASTRO radiotherapy clinic in Maastricht sees various types of cancer patients and treats them with radiotherapy. MAASTRO clinic recruits eligible patients for various clinical trials. Research nurses pre-screen incoming patients to see if patients are potentially eligible to participate in a clinical trial. If a patient is potentially eligible then the clinical trial and standard treatment are discussed with the patient at intake.

Several breast cancer clinical trials that are or were previously running at MAASTRO will be encoded in the trial metadata repository, as well as several clinical trials from the German Breast Group (GBG). These trials have a mix of trials that were unfamiliar and familiar to the trial nurse at MAASTRO. Further this could simulate both monitoring and recruiting for ongoing trials running on site, as well as the moment a site either is considering joining a trial or has just begun to recruit for a trial. These trials were run against an anonymized breast cancer patient data set provided by GBG with 4673 patients.

The evaluation will follow the following protocol:

10 min	Introduction • Welcome • Consent Form • Intro to who we are
10-20 min	Interview • Brief intro to trial recruitment • Discuss background of participant related to trial recruitment and their current work situation
10-15 min	Training • Demonstrate and give a walkthrough of Yakobo to the participant • During this time researcher will answer any questions participant has • Use one or two test trials for the training
15-20min	Actual Usage • Participant reads the scenario • Participant inspects and uses for one or two trials
10-20 min	Discussion • Anything usage of Yakobo prompts the participant to discuss • Discuss applications of Yakobo in their work
5 min	Questionnaires and Debriefing
Total = 60-90 min	

Patients are referred to the MAASTRO radiotherapy clinic by various regional hospitals in Limburg when radiotherapy is prescribed for their cancer treatment. Research nurses pre-screen patients for eligibility in clinical trials in the days before a patient arrives for intake at the MAASTRO clinic. There are potentially multiple trials that a patient could be enrolled in per disease type. The trials are prioritized according to an internal agreement. A patient is screened for these trials, checking trials in the order of priority. If a patient is found to be potentially eligible for a trial, that treatment option and standard treatment are discussed with the patient at intake.

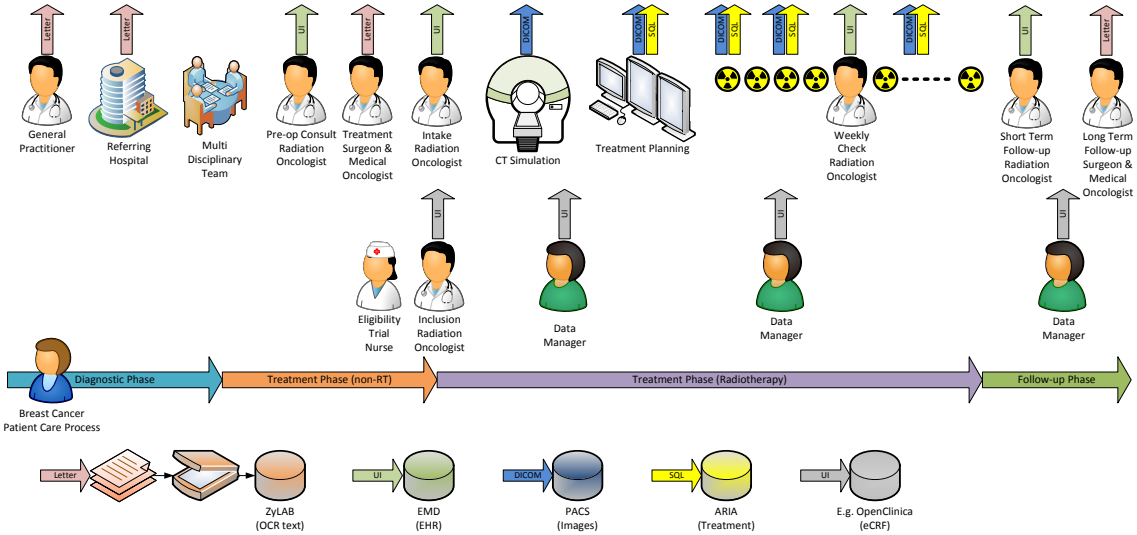


Figure 7: MAASTRO care process and data sources. Top: Radiotherapy care process including data generation (arrows). Top-Middle: Additional data generation if patient is included in a clinical trial. Bottom-Middle: Chronological treatment phases (not to scale).

The MAASTRO clinic is invited to participate in both national, and particularly international clinical trials. Investigators from other sites talk with investigators from MAASTRO when they want to expand their population, and ask whether MAASTRO would like to join the trial. Before active recruiting can begin, there is a check to see whether there is a relevant population of patients arriving at MAASTRO. This is done currently using the general statistics that are collected internally (e.g., they know the types of tumors they irradiate at the clinic, the number of patients per year they see etc.) and some statistical analysis. Once this check has been done, and MAASTRO is considered a good site to participate, the relevant setup work is done, and then the clinical trial becomes one of the options in the regular pre-screening process. The MAASTRO care process and the relevant data sources are presented in Figure 7.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 5: Main features of the evaluation procedure for the "Trial recruitment" scenario.

		Description
General	Evaluation Scenario	Trial Recruitment
	Evaluation Leader	Maastro

	Description of the scenario		Eligible patients are identified on a trial-first basis in a clinic where research nurses usually identify trials on a patient-first basis.
	Description of the evaluation procedure		The user is presented with a trial and eligible patients for that trial. The user can then look at the eligibility criteria that are known for each patient to evaluate correct eligibility.
Setup	Location of evaluation		MAASTRO
	Setup details		The system can be used from a laptop provided by Philips.
	Evaluator's expertise		Research nurse for radiotherapy trials
	Number of evaluators		2
	Duration of the evaluation		2 hours
	Time plan		End of April 2015
Data	Kind of data used for the evaluation		Retrospective
	For retrospective data		Patient data from MAASTRO and German Breast Group
	If you plan to access CDP servers		Yes
	If you plan on using prospective data		N/A
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		A trial-first tool for the identification of eligible patients
	How it will be evaluated?		A MAASTRO Research Nurse will use the interface to select patients for familiar trials from MAASTRO and unfamiliar trials from German Breast Group.
	How many clinicians/users will be involved?		One
	How many times will they fill-in the usability questionnaires?		Once
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Sequential
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	N/A

		Inclusion/Exclusion Criteria?	As defined by the selected trials and prepared patient data
		Will the EURECA tool be compared with the tool you normally use?	N/A
		Which is the evaluation hypothesis?	The research nurse will find that the tool simplifies and speeds up the task of identifying patients eligible for a trial.
	Evaluation factor 1	Quality metrics	Usability
		Measured value	Score from Usability Questionnaire
		Rating levels	Score 1 - 100
	Criteria	Assessment criteria	If our minimum acceptance level is met, the tool will be likely useful for continued use as a graphical user interface to another tool under development.

6 Protocol feasibility

Protocol feasibility is a process that assesses the feasibility of a clinical protocol under development. The main goal is to eliminate possible errors in the protocol description in an early stage of the development and find an optimal recruitment population of patients (reducing costs). Central aim of the specific scenario is the creation of eligibility criteria in the protocol. These eligibility criteria will be central created/formalised in the protocol feasibility tool and send to local hospital sites for validation. The sites will return aggregated patient results for the given criteria, which will give the study managers an idea of feasibility.
The web front-end will be evaluated by a selected set of study managers and others with experience with protocol feasibility.

The evaluation will be done in sessions, structured which are structured like this:
- Introduction of tool (presentation by developer)
- Let the user work with the tool
- Formal feedback from end-user (questionnaires)
- Informal feedback from end-user

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 6: Main features of the evaluation procedure for the "Protocol feasibility" scenario.

		Description
General	Evaluation Scenario	Distributed Protocol Feasibility
	Evaluation Leader	GBG - Custodix - Maastro
	Description of the scenario	Eligibility criteria will be central created/formalized in the protocol feasibility tool and send to local hospital sites for validation. The sites will return aggregated patient results for the given criteria, which will give the study managers an idea of feasibility.
	Description of the evaluation procedure	The evaluation will be done in sessions, structured which are structured like this: • Introduction of tool (presentation by developer) • Let the user work with the tool • Formal feedback from end-user (questionnaires) • Informal feedback from end-user
Setup	Location of evaluation	Evaluation sessions at Frankfurt hospital
	Setup details	local site deployments (semantic layer, local feasibility endpoints) at GBG and Maastro and centralized deployment (central feasibility tool)
	Evaluator's expertise	Study managers and physicians
	Number of evaluators	One
	Duration of the evaluation	2 hour sessions

	Time plan	1. Introduction (5 min) 2. Interview (20 min) 3. Training (25 min) 4. Discussion (20 min) 5. Metrics (25 min) 6. Questionnaires (5 min) 7. Debriefing (5 min) End of May 2015
Data	Kind of data used for the evaluation	prospective
	For retrospective data	N/A
	If you plan to access CDP servers	Local deployment of the semantic layer (data at site)
	If you plan on using prospective data	(depends on site)
Evaluation type	Type of evaluation	Proof of concept
	Set-up (only for Clinical testing)	Retrospective
Evaluation factor 1	Quality metrics	• Correctness • Timing • Usability • The outcome of the questionnaires A and B
	Measured value	Not determined yet
	Rating levels	Not determined yet
Criteria	Assessment criteria	SUS score of the questionnaires

7 Update of guidelines

The scenario of guideline update provides the service for professionals (clinicians, researchers, and guideline designers) to find new relevant evidence from biomedical search engines to check whether or not a selected guideline statement of a medical guideline needs to be updated.
The evaluation will take place in March 2015 and be carried out by MAASTRO doctors familiar or charged with updating the national guidelines for radiotherapy treatment of NSCLC (Non-Small Cell Lung Carcinoma). The evaluation procedure of guideline update consists of the following steps: i) by selecting a medical guideline, the system will return a set of the guideline topics (namely section titles of guideline); ii) by selecting a topic of the guideline, the system will return a set of guideline statements with their supported evidence; iii) by selecting a guideline statement and click on the button "Finding relevant evidence", the system will search the PubMed site and return a set of articles with the estimated evidence quality; iv) the user will check which newly discovered articles are relevant to the update of the selected guideline statement.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 7: Main features of the evaluation procedure for the "Update of guidelines" scenario.

		Description
General	Evaluation Scenario	Update of guidelines
	Evaluation Leader	MAASTRO
	Description of the scenario	The scenario of guideline update provides the service for professionals (clinicians, researchers, and guideline designers) to find new relevant evidence from biomedical search engines to check whether or not a selected guideline statement of a medical guideline needs to be updated.
	Description of the evaluation procedure	The evaluation procedure of guideline update consists of the following steps: i) by selecting a medical guideline, the system will return a set of the guideline topics (namely section titles of guideline); ii) by selecting a topic of the guideline, the system will return a set of guideline statements with their supported evidence; iii) by selecting a guideline statement and click on the button "Finding relevant evidences", the system will search the PubMed site and return a set of articles with the estimated evidence quality; iv) the user will check which newly discovered articles are relevant to the update of the selected guideline statement.
Setup	Location of evaluation	MAASTRO Clinic

	Setup details		The system can be installed on the laptop of the users.
	Evaluator's expertise		Physicians, researchers, or guideline designers
	Number of evaluators		3
	Duration of the evaluation		One week
	Time plan		End of April 2015
Data	Kind of data used for the evaluation		Prospective
	For retrospective data		N/A
	If you plan to access CDP servers		N/A
	If you plan on using prospective data		N/A
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		A tool for the update of medical guidelines
	How it will be evaluated?		A MAASTRO employee who works on the national committee that develops guidelines for radiotherapy treatment of NSCLC lung cancer will use the Guideline Update tool to check against a recent guideline update.
	How many clinicians/users will be involved?		One
	How many times will they fill-in the usability questionnaires?		Once (possibly twice after taking initial feedback into account).
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Sequential
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	N/A
		Inclusion/Exclusion Criteria?	N/A
		Will the EURECA tool be compared with the tool you normally use?	N/A
		Which is the evaluation hypothesis?	

	Evaluation factor 1	Quality metrics	How many articles are relevant to the guideline update.
		Measured value	The number of relevant articles found.
		Rating levels	Minimal acceptance: for half of selected guideline statements, at least one relevant evidence can be found.
	Evaluation factor 2	Quality metrics	Percentage of discovered articles deemed to be relevant.
		Measured value	
		Rating levels	
	Criteria	Assessment criteria	If our minimum acceptance level is met, the tool will be likely useful for continued use for the tested guideline (radiotherapy treatment of NSCLC lung cancer). The number of relevant articles found should be as small as possible, ideally less than 20.

8 Hypothesis generation

Hypothesis generation is a tool that allows clinicians to generate and evaluate hypotheses in the context of designing a clinical trial. Data are uploaded via KDF functionality. Scripts are uploaded via KDF. Jobs are managed via KDF. The evaluation will consist of interviews with the users, who will be presented with the tool and a questionnaire to assess the usefulness of this tool.

The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool.

The evaluation will start with a demonstration that aims to provide guidance for the evaluators and followed by a number of tasks to complete. The following steps detail evaluation plan of the EURECA tool.

1. Discuss the possible evaluation dates and provide background information for potential evaluators.
2. Install the KDF on the evaluation site, which needs to be decided with the evaluators.
3. In the first part of the evaluation, the evaluators will be given a demo about the system. Particularly the staff will show the users how to select datasets, upload the script and run the data mining job.
4. In the second part, at first the evaluators will perform the same task but using different scripts. After that the results will be checked and interpreted by the evaluators. A set of measurements will be made, which are listed as evaluation factors. These heuristic evaluations are based on literature [3].
5. Then the evaluators will be asked to write their own script that can be uploaded to the KDF platform and run the data mining job. Similarly the results need to be interpreted.
6. After the exercise, the evaluators will answer the evaluation questionnaire.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 8: Main features of the evaluation procedure for the "Hypothesis generation" scenario.

		Description
General	Evaluation Scenario	Hypothesis generation
	Evaluation Leader	UOXF
	Description of the scenario	A tool that allows clinicians to generate and evaluate hypotheses in the context of designing a clinical trial. Data are uploaded via KDF functionality. Scripts are uploaded via KDF. Jobs are managed via KDF. The evaluation will consist of interviews with the users, who will be presented with the tool and a questionnaire to assess the

		usefulness of this tool.
	Description of the evaluation procedure	The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool.
Setup	Location of evaluation	Oncology department of Oxford University
	Setup details	The tool will be installed locally and will be available only to the evaluator.
	Evaluator's expertise	A biostatistician or bioinformatician
	Number of evaluators	3
	Duration of the evaluation	Maximum 1 hour
	Time plan	When the latest KDF is updated (Last week of April 2015)
Data	Kind of data used for the evaluation	Retrospective
	For retrospective data	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).
	If you plan to access CDP servers	(a)Yes (b)Yes
	If you plan on using prospective data	Not planned
	Type(s) of evaluation	Proof of concept
Proof of Concept	What exactly will be evaluated?	The evaluation will test the validity of the KDF and the usability of the outcome prediction functionality.
	How it will be evaluated?	The evaluator will load the clinical data and upload the scripts using the KDF and the usability and functionality will be assessed by answering the questionnaire.
	How many clinicians/users will be involved?	Three
	How many times will they fill-in the usability questionnaires?	Once

Clinical Evaluation	Set-Up	Clinical Evaluation Type	Same end-users will do the evaluation. It will take place at the same location with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Clinical information of 219 patients, all consented and published data
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	None
		Which is the evaluation hypothesis?	With the EURECA Hypothesis generation tool it is possible to identify new hypothesis for further studies.
	Evaluation factor 1	Quality metrics	Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc.
		Measured value	Time
		Rating levels	
	Evaluation factor 2	Quality metrics	The number of alternative paths that analysts are able to explore, i.e. with the help of the tool how many possible ways can be thought of to generate different hypotheses by the same subject.
		Measured value	Integer
		Rating levels	
	Evaluation factor 3	Quality metrics	The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task
		Measured value	Integer
		Rating levels	
	Criteria	Assessment criteria	Usability assessment based on ISO/IEC 25000 series and on SUS scale

9 Outcome prediction

The outcome prediction tool allows clinicians or researcher to predict the outcome of treatment for a patient. Data are uploaded via KDF functionality. Script are uploaded via KDF. Jobs are managed via KDF.

The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool. The evaluation will follow the same plan as described for the Hypothesis generation scenario.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 9: Main features of the evaluation procedure for the "Outcome prediction" scenario.

		Description
General	Evaluation Scenario	Outcome prediction
	Evaluation Leader	UOXF
	Description of the scenario	A tool that allows clinicians or researcher to predict the outcome of treatment for a patient. Data are uploaded via KDF functionality. Script are uploaded via KDF. Jobs are managed via KDF.
	Description of the evaluation procedure	The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool.
Setup	Location of evaluation	Oncology department of Oxford University
	Setup details	The tool will be installed locally and will be available only to the evaluator.
	Evaluator's expertise	A biostatistician or bioinformatician
	Number of evaluators	3
	Duration of the evaluation	Maximum 1 hour
	Time plan	When the latest KDF is updated (Last week of April 2015)
Data	Kind of data used for the evaluation	Retrospective
	For retrospective data	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).
	If you plan to access CDP servers	(a)Yes (b)Yes

	If you plan on using prospective data		Not planned
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		The evaluation will test the validity of the KDF and the usability of the outcome prediction functionality.
	How it will be evaluated?		The evaluator will load the clinical data and upload the scripts using the KDF and the usability and functionality will be assessed by answering the questionnaire.
	How many clinicians/users will be involved?		Three
	How many times will they fill-in the usability questionnaires?		Once
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation. It will take place at the same location with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Clinical information of 219 patients, all consented and published data.
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	None
		Which is the evaluation hypothesis?	With the EURECA outcome prediction tool it is possible to create predictive models from different clinical sources.
	Evaluation factor 1	Quality metrics	Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc.
		Measured value	Time
		Rating levels	

	Evaluation factor 2	Quality metrics	The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task.
		Measured value	Integer
		Rating levels	
	Evaluation factor 3	Quality metrics	Performance: • For prediction of binary outcome: AUC = Area under the ROC (receiver operating characteristic) curve • For prediction of survival outcome: c-index = concordance index, a performance measure specifically for censored data
		Measured value	AUC (data type: double) or c-index (data type: double)
		Rating levels	
	Criteria	Assessment criteria	Usability assessment based on ISO/IEC 25000 series and on SUS scale

10 Diagnostic classifier

The diagnostic classifier tool allows clinicians or researcher to classify patients into diagnostic groups. Data are uploaded via KDF functionality. Script are uploaded via KDF. Jobs are managed via KDF.

The evaluation will consist of interviews with the users, who will be presented with the tool and a questionnaire to assess the usefulness of this tool. The evaluation will follow the same plan as described for the Hypothesis generation scenario. The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 10: Main feature of the evaluation procedure for the "Diagnostic classifier" scenario.

		Description
General	Evaluation Scenario	Diagnostic classifier
	Evaluation Leader	UOXF
	Description of the scenario	A tool that allows clinicians or researcher to classify patients into diagnostic groups. Data are uploaded via KDF functionality. Scripts are uploaded via KDF. Jobs are managed via KDF. The evaluation will consist of interviews with the users, who will be presented with the tool and a questionnaire to assess the usefulness of this tool.
	Description of the evaluation procedure	The KDF will be installed at the evaluation site and a biostatistician will use the tool to predict the outcome based on clinical data of patients and fill out an evaluation questionnaire about the tool.
Setup	Location of evaluation	Oncology department of Oxford University
	Setup details	The tool will be installed locally and will be available only to the evaluator.
	Evaluator's expertise	A biostatistician or bioinformatician
	Number of evaluators	
	Duration of the evaluation	Maximum 1 hour
	Time plan	When the latest KDF is updated (Last week of April 2015)
Data	Kind of data used for the evaluation	Retrospective
	For retrospective data	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).

	If you plan to access CDP servers		(a)Yes (b)Yes
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		The evaluation will test the validity of the KDF and the usability of the outcome prediction functionality.
	How it will be evaluated?		The evaluator will load the clinical data and upload the scripts using the KDF and the usability and functionality will be assessed by answering the questionnaire.
	How many clinicians/users will be involved?		Three
	How many times will they fill-in the usability questionnaires?		Once
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation. It will take place at the same location with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Clinical information of 219 patients, all consented and published data
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	None
		Which is the evaluation hypothesis?	With the EURECA diagnostic classifier tool it is possible to create predictive models from different clinical sources.
	Evaluation factor 1	Quality metrics	Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc.
		Measured value	Time
		Rating levels	

	Evaluation factor 2	Quality metrics	The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task.
		Measured value	Integer
		Rating levels	
	Evaluation factor 3	Quality metrics	Robustness of clustering: a similarity measure for the patient clusters, if cross-validation or bootstrapping procedures are applied.
		Measured value	Robustness (data type: double)
		Rating levels	
	Criteria	Assessment criteria	Usability assessment based on ISO/IEC 25000 series and on SUS scale

11 Prediction of SAEs/SUSARs

A study protocol describes, which AE criteria should be reported as a SAE (or SUSAR). These criteria are evaluated by the Ethics Committee before the trial starts. In some cases these criteria have to be changed afterwards. If a toxicity table (e.g. CTCAE) is included in the protocol, AE terms and the grading scale provided should be used. SAEs and SUSARs of a trial will be (automatically) reported through a specific CRF in the CT system.

These reports can be used in an analysis in order to try to find predicting variables for the SAE. These predicting variables can then be used to generate a risk score for each new patient. Based on that risk score, decisions can be made regarding e.g. choice of treatment, chemotherapy dosage or frequency of monitoring the patient.

Initially, there was a request from the University of Saarland to create a prediction model for an SAE from data collected in a trial. This data consisted of high-level features such as age, weight, tumor location and details of the treatment regime, no genetic or imaging data was included. This dataset consisted of information from 3120 patients with Wilms' tumor, or nephroblastoma, enrolled in the SIOP trial [1]. Wilms' tumor, is a kidney cancer which occurs mainly in children younger than five years of age [2]. Treatment consists of preoperative chemotherapy, tumornephrectomy (surgical removal of the kidney with the tumor) and postoperative chemotherapy and sometimes radiotherapy according to histology and stage of the tumor [3].

In the SIOP Nephroblastoma study several types of information have been recorded via Case Report Forms (CRFs, forms used in clinical trials to record data), including SAEs, consultation data, surgery complications, measured volume reduction of the tumor after pre-operative chemotherapy and relapses, deaths and survival rate (the follow-up lasted for five-years [3]). The University of Saarland is interested in finding out whether we can predict from patient characteristics and history and the specifics of the treatment (e.g. dosage of chemotherapy), whether a patient will develop Veno-Occlusive Disease (VOD) in response to treatment.

The features that were selected from the data set are:

- the patient's age at start of treatment
- the patient's weight at start of treatment
- the location of the tumor (left, right, bilateral or extrarenal)
- the dosage of radiotherapy at each of the above mentioned locations
- the dosage of preoperative chemotherapy during a maximum of 24 weeks (in case of bilateral disease) before surgery
- the dosage of postoperative chemotherapy during a maximum of 35 weeks after surgery

Occurrences of VOD were recorded in CRF's reporting on the preoperative and postoperative chemotherapy phase of treatment with a time stamp relative to start of treatment.

During discussions regarding the details of the envisioned prediction model and the details of the supplied data set, it became clear that generation of such a prediction model is a very complex task, requiring cross-expertise communication; the clinical

expert needs to communicate his/her knowledge regarding the specifics of the recorded data to the data mining expert and the data mining expert needs to convey his/her needs regarding the prediction modeling proces to the clinical expert. This is a very time consuming and difficult process. In order to facilitate this process of communication, a tool is envisioned which provides a user interface that can be understood by the clinical expert and can be used by this clinical expert together with the data mining expert to construct a prediction model quickly, so that the clinical expert can easily convey the required knowledge of the data set to the data mining expert, who can then use this crude first version of the prediction model, to tease out the details and see if it can be improved.

Such a tool would exploit the EURECA common data model and the tools and services developed around it, so that other prediction modeling scenarios can make use of it in a similar manner. In order to demonstrate the genericity of the tool, a second scenario was added: MAASTRO clinic will supply a data set containing data from lung cancer patients who have been treated with radiation therapy at the MAASTRO clinic. This data set has recorded two SAEs for which there is interest from MAASTRO clinic to build prediction models, namely Dyspnea and Dysphagia. Prediction models for the UdS as well as the MAASTRO use case will be constructed by physicians from both sites as part of the evaluation.

The evaluation will follow the following protocol:

10 min	Introduction • Welcome • Consent Form • Intro to who we are
10-20 min	Interview • Brief intro to prediction modeling • Discuss background of participant related prediction modeling and their current work situation
10-15 min	Training • Demonstrate and give a walkthrough of SAE prediction tool to the participant • During this time researcher will answer any questions participant has • Use one or two test scenarios for the training
15-20min	Actual Usage • Participant reads the scenario • Participant uses SAE prediction tool to generate prediction models for the scenario and inspects the result (discussing it with the researcher)
10-20 min	Discussion • Anything usage of the SAE prediction tool prompts the participant to discuss • Discuss applications of the tool in their work
5 min	Questionnaires and Debriefing
Total = 60-90 min	

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 11: Main features of the evaluation procedure for the "Prediction of SAEs/SUSARs" scenario.

		Description
General	Evaluation Scenario	Prediction of SAE
	Evaluation Leader	Norbert Graf
	Description of the scenario	A study protocol describes, which AE criteria should be reported as a SAE (or SUSAR). These criteria are evaluated by the Ethics Committee before the trial starts. In some cases these criteria have to be changed afterwards. If a toxicity table (e.g. CTCAE) is included in the protocol, AE terms and the grading scale provided should be used. SAEs and SUSARs of a trial will be (automatically) reported through a specific CRF in the CT system.
	Description of the evaluation procedure	Both from UdS and from MAASTRO, a physician and a data mining expert are invited to execute the scenarios that have been submitted by their sites. For each site, a couple consisting of a physician and a data mining expert who have worked together to create prediction models in the past are invited. The usage of the tool is explained to them and they are asked to perform a practice scenario together with the instructor. After they have completed this practice scenario, and have had time to ask any remaining questions regarding the tool, they are asked to perform their own site's scenario together and to narrate the steps that they are taking. The evaluation procedure is concluded with a discussion comparing the previous way of working of this couple with the way of working supported by the tool. A usability questionnaire is also given to both users and they are asked to fill in this questionnaire separately.
Setup	Location of evaluation	UdS - Hospital for Paediatric Oncology and Haematology
	Setup details	Local setup of the SAE prediction tool frontend as well as the python backend (running as python webserver), connection to the CDM remote (set up via the SAE prediction tool frontend, which contains a login screen).

	Evaluator's expertise		Professor for Paediatric Oncology Haematology
	Number of evaluators		At least 3 physicians and 3 study nurses
	Duration of the evaluation		This will be done within 4 weeks after setup of the system, duration of the evaluation will be 60-90 minutes per evaluator.
	Time plan		Beginning of May 2015 after the installation of the system.
Data	Kind of data used for the evaluation		Retrospective or fake data will be used
	For retrospective data		SIOP 2001 data, if the prediction will be done for VOD, accessed through the CDP servers.
	If you plan to access CDP servers		Not planned
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		The evaluation will be a proof of concept but also asking the question, if the tool can be used in clinical care.
Proof of Concept	What exactly will be evaluated?		- Usability of the SAE prediction tool by couples of clinical and data mining experts with the goal to create prediction models more quickly and with a better understanding of the proces by both types of users. - Accuracy of the prediction models that are generated
	How it will be evaluated?		The templates given in D8.1. will be used.
	How many clinicians/users will be involved?		At least 3 clinicians and 3 study nurses
	How many times will they fill-in the usability questionnaires?		If the tool is not satisfying the end-users there will be an iterative process with the developers to optimize the tool and a new evaluation will be done by the same end-users.
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation in an iterative process as long as the tool is satisfying all end-users. It will take place at the same location (USAAR) with the same parameters.

		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	SIOP 2001 data, if the prediction will be done for VOD. 3 physicians and 3 study nurses.
		Inclusion/Exclusion Criteria?	
		Will the EURECA tool be compared with the tool you normally use?	There is no tool that is normally used
		Which is the evaluation hypothesis?	Using the SAE Prediction tool to collaborate in order to generate prediction models facilitates quicker mutual understanding between clinician and data mining expert, thereby facilitating: - the data mining expert in obtaining the right requirements from the clinician and the knowledge about the data required for the creation of the prediction model and - the clinician in obtaining a better understanding of the proces which lead to the prediction model, thereby creating a higher level of trust in the final model teased out by the data mining expert based on the first version resulting from the collaboration using the tool.
	Evaluation factor 1	Quality metrics	Usability of the tool and accuracy of the prediction models
		Measured value	Usability of the tool and accuracy of the prediction models
		Rating levels	The rating given in D8.1 is used.
	Criteria	Assessment criteria	Assessment criteria are the usability during clinical practice and the percentage of predicted SAEs

12 Patient Diary & Long-term follow-up

Evaluation will be held inside the NHS UK. FORTH will provide a fully functional virtual machine with the PHR and the EURECA extensions. Real patient data will be imported to the system (bulk import). Clinicians from NHS will use & evaluate the tool.

Evaluation factors/parameters:

- Survey PHR system evaluation criteria/questionnaire
- Design interview with clinicians who evaluate the PHR system
- Answer the question:
  - Would you recommend this tool to your patients?

The flow of operations is shown in the following figure.

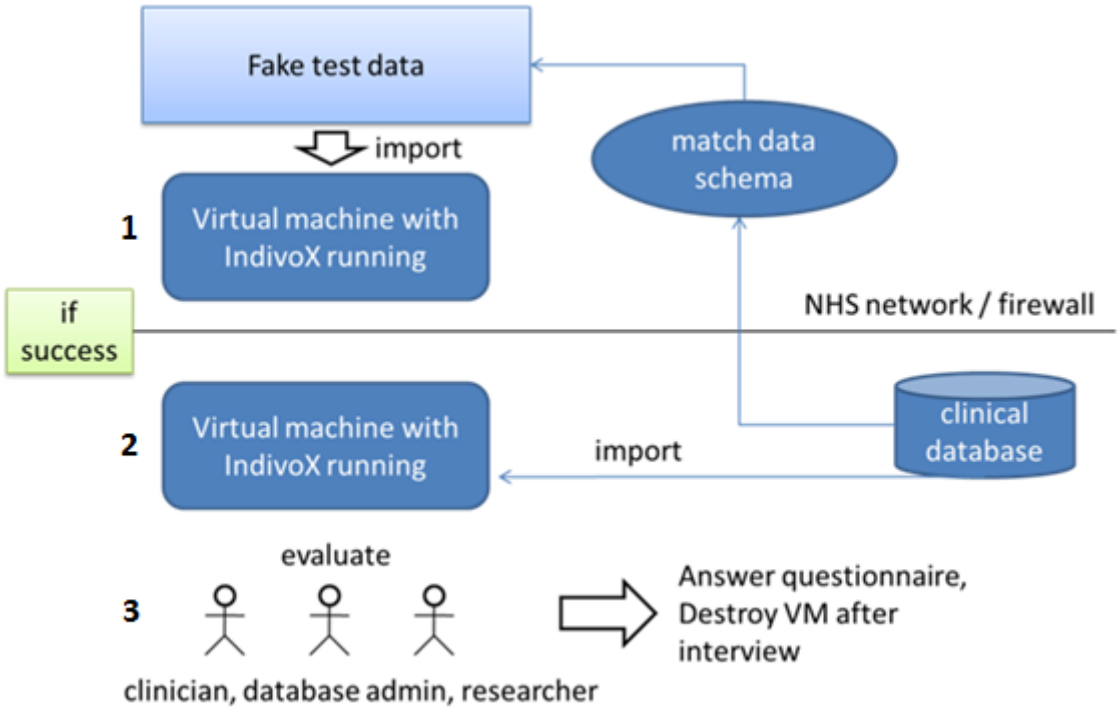


Figure 8: The flow operation for the evaluation of the "Patient diary & long-term follow-up" scenario.

Data and the virtual machine will be destroyed after the interview.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 12: The main features of the evaluation procedure for the "Patient diary & long-term follow-up" scenario.

		Description
General	Evaluation Scenario	Patient diary & Long-term follow-up
	Evaluation Leader	UOXF
	Description of the scenario	A patient health record system (IndivoX with EURECA extensions) that records patient information such as medication, procedures, lab results etc.
	Description of the evaluation procedure	Evaluation will be held inside the NHS UK. FORTH will provide a fully functional virtual machine with the PHR and the EURECA extensions. Real patient data will be imported to the system (bulk import). Clinicians from NHS will use & evaluate the tool.
Setup	Location of evaluation	Oncology department of Oxford University
	Setup details	The tool will be installed locally and only available to the evaluator
	Evaluator's expertise	Clinician / Doctor
	Number of evaluators	3 or 4
	Duration of the evaluation	Half an hour
	Time plan	End of April, 2015
Data	Kind of data used for the evaluation	Retrospective
	For retrospective data	Access the CDP servers to perform the validation (the server of UPM, or of Custodix)
	If you plan to access CDP servers	(a)Yes (b)Yes
	If you plan on using prospective data	Not planned
	Type(s) of evaluation	Proof of concept
Proof of Concept	What exactly will be evaluated?	The functionality and the usability of the PHR system will be evaluated by the clinicians.
	How it will be evaluated?	The clinicians will have 30 minutes to evaluate the tool. In the first 10 minutes the evaluators will see a short demonstration about the PHR system and the functionalities. Then in the other 10 minutes they will be guided to use the tool, e.g. viewing patient information, adding new data, etc. After that, they will answer the questions on the questionnaire and comment on the system in the remaining time.

	How many clinicians/users will be involved?		Three
	How many times will they fill-in the usability questionnaires?		Once
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation. It will take place at the same location with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	Clinical information of NHS patients, local data that never leave the NHS network.
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	None
		Which is the evaluation hypothesis?	With the EURECA PHR tool it is easy to view patient diary data and provide an easy way to interact with patient data.
	Evaluation factor 1	Quality metrics	Time to create a new event, or a new appointment on the calendar.
		Measured value	Time
		Rating levels	
	Evaluation factor 2	Quality metrics	Time of inserting, editing deleting a procedure and a medication.
		Measured value	Time
		Rating levels	
	Evaluation factor 3	Quality metrics	Accuracy of the drug-drug interaction
		Measured value	Percentage of error
		Rating levels	

	Criteria	Assessment criteria	Usability assessment based on ISO/IEC 25000 series and on SUS scale

13 Automatic detection and reporting of SAEs/SUSARs

The reporting service is integrated in ObTiMA, a multifunctional clinical trial management tool. It is built up on already existing functionalities. After the occurrence of an adverse event authorized clinical trial personnel has to complete predefined safety documentation forms (AE, SAE CRFs). The trial sponsor will be automatically informed about the safety issue. After verification and assessment of causality and expectedness by the trial sponsor, a SUSAR Fax will be generated. In this case additional entries will be linked to the service related forms to ensure that the SUSAR fax meet the EMA pharmacovigilance requirements [4] [5] [6]. This procedure is visualised in Figure 9.

Below the SAE/ SUSAR procedure is summarized:

- *AE/ SAE data entry (investigator)*
- *AE/ SAE assessment (investigator)*
- *AE/ SAE verification and evaluation (sponsor/ trial chairman)*
- *SAE/SUSAR confirmation and notification to the concerned competent authorities, ethic committee(s)and investigators (sponsor/ trial chairman)*

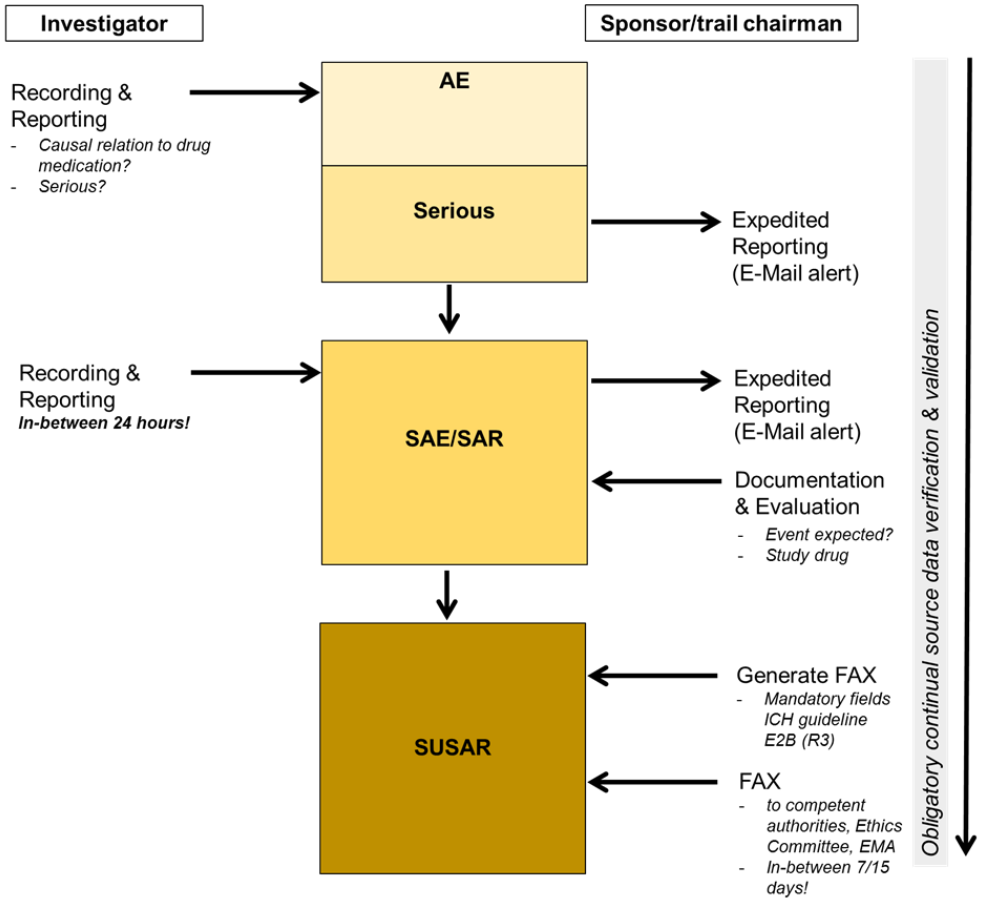


Figure 9: General SAE/ SUSAR reporting procedure

The SAE/SUSAR Reporting Service is integrated in the Web Application ObTiMA. To test the SAE/ SUSAR Reporting Service an Internet Browser and an Internet Connection are required.
The evaluation of the features of the SAE/ SUSAR Reporting Service from the investigator's perspective is described. Please follow the points step by step. In summary, these steps are:

1. Open the ObTiMA web application.
2. Open the AE form in the documentation of the concerned patient.
3. A) Enter event specific patient data, provide a causality assessment and mark the event as 'serious'. The sponsor will be automatically alerted via E-Mail to the occurrence of a new SAE.

   B) Check the receipt of the alert message.
4. Complete SAE form. The sponsor will be automatically alerted via E-Mail to the completion of the SAE form.

Log in

You have to log-in in the ObTiMA application first.

Open patient specific AE Form

The AE Form is deposited in the respective trial patient specific eCRF.

1. Select the clinical trial in the trials menu (Figure 10).

Figure 10: Clinical trial selection in ObTiMA

2. Select the patient, who is/has been affected by an (serious) adverse event (Figure 11).

Figure 11: Selection of the affected patient

3. Open the AE form which is located in the respective study event (Figure 12).

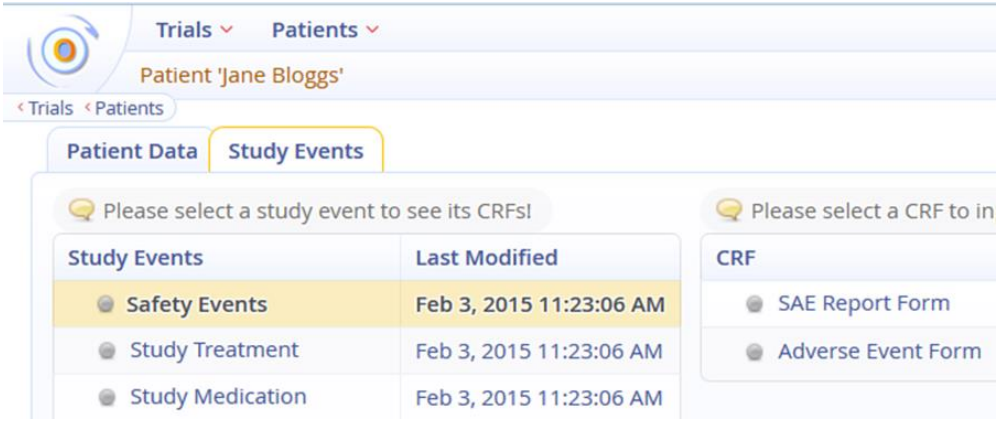


Figure 12: Patient Folder Overview

AE Data Entry

1. Complete the AE form with all available data from the patient's medical records.
2. Provide a causality assessment and mark the event as 'serious' (Figure 13).
3. Save your AE CRF entries. An alert message will automatically be sent to the sponsor/ trial chairman and you, furthermore a SAE CRF opens automatically.

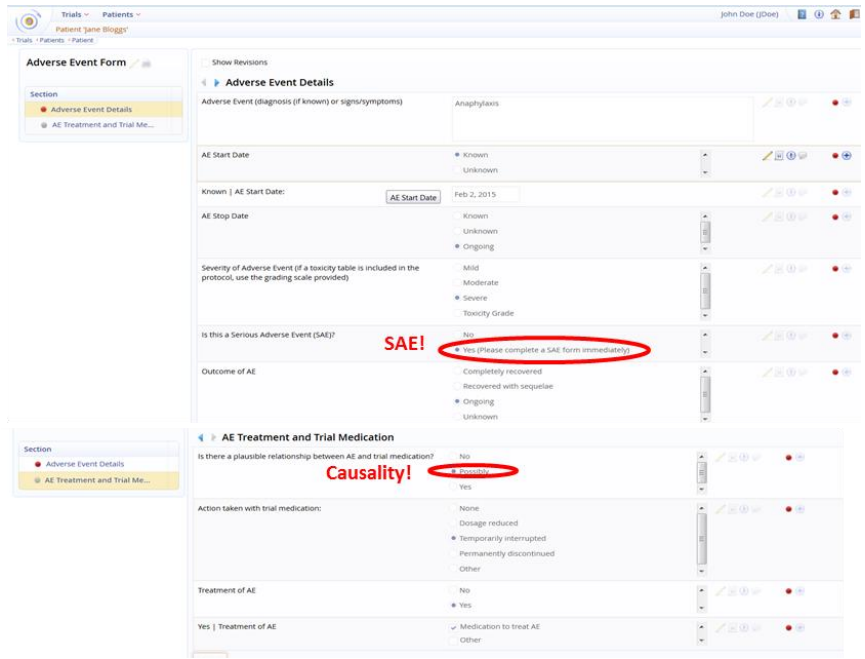


Figure 13: AE data entry

SAE Data entry

1. Enter all available data from the patient's medical records in the SAE CRF within 24 hours (Figure 14).
2. All mandatory data fields (marked with ✱) have to be completed; otherwise the SAE CRF cannot be saved.
3. After you have saved the SAE CRF an alert message will automatically be sent to the sponsor/ trial chairman and to you.

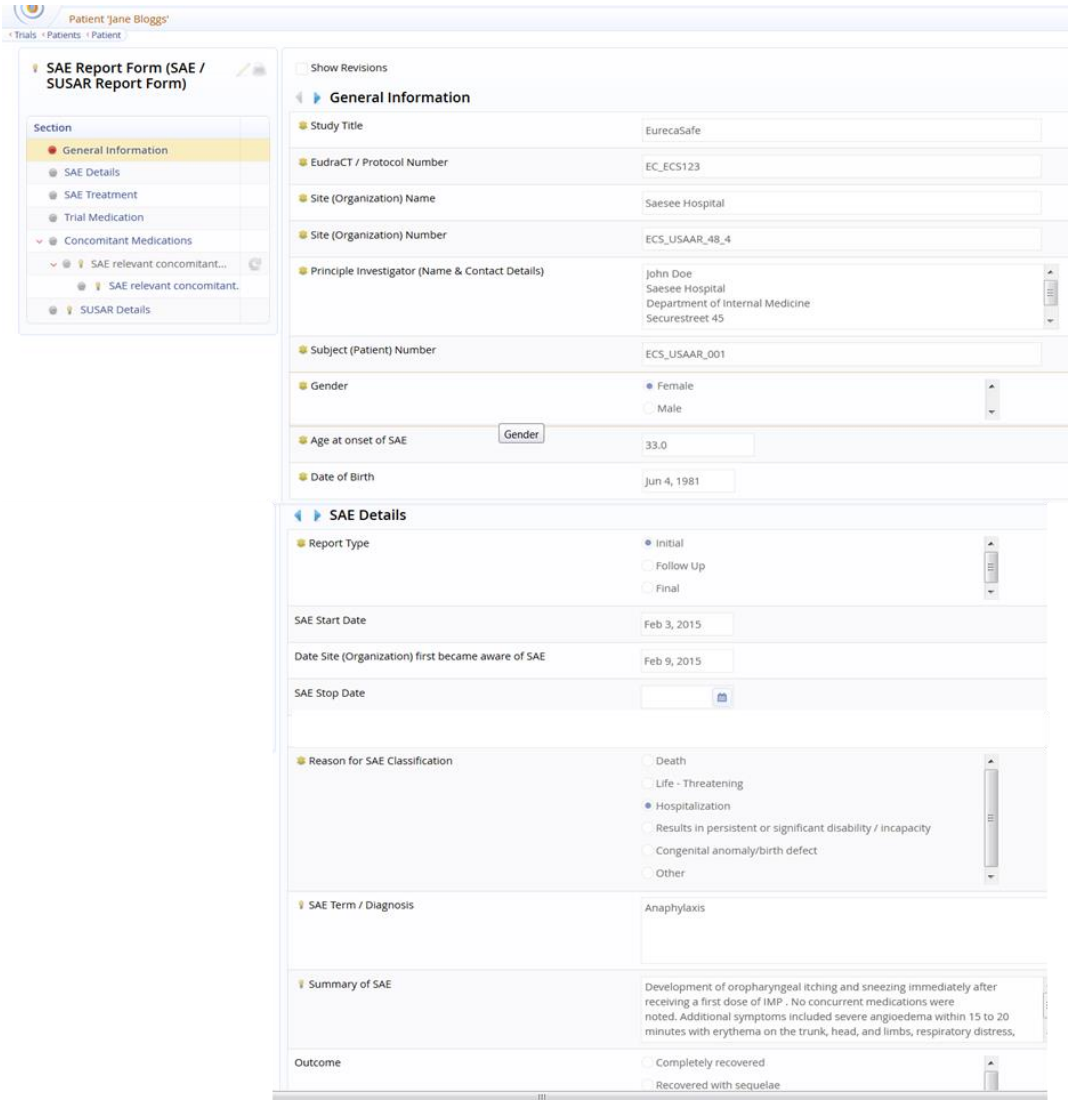


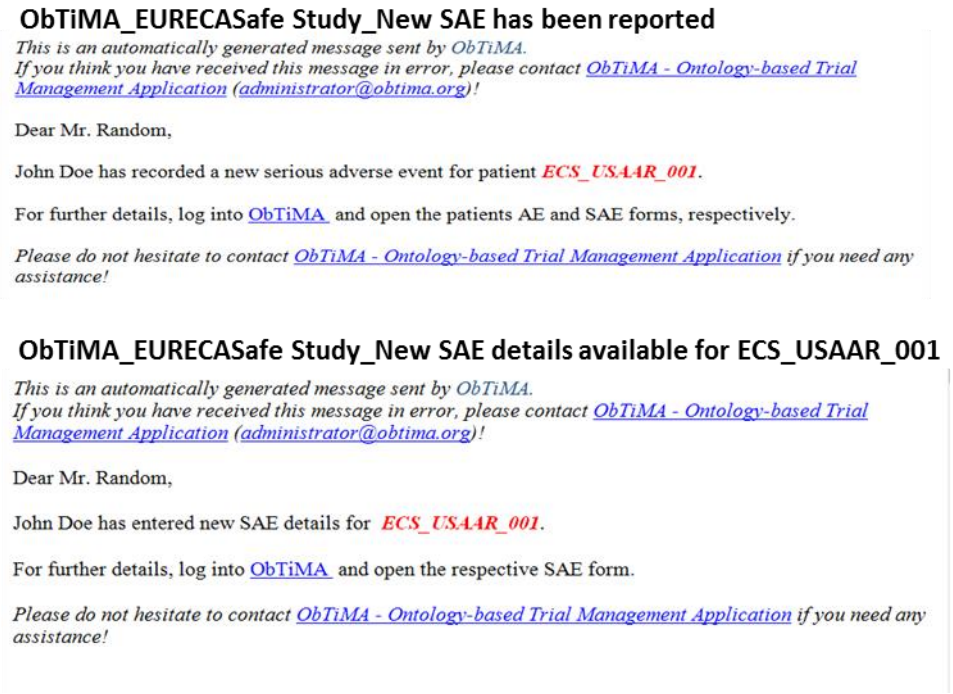
Figure 14: SAE Data Entry

Evaluation Scenario sponsor/ trial chairman

In the following the evaluation of the features of the SAE/ SUSAR reporting service from the sponsor's/ trial chairman's perspective is described. Please follow the points step by step. In summary, these steps are:

1. Check your E-Mail for incoming safety alerts (Figure 15).
2. Open the respective AE/ SAE form in the documentation of the concerned patient.
3. Validate (source data verification/ medical review/ coding) the entered data.
4. After assessment of the causality and the expectedness complete the sponsor/ trial chairman specific mandatory fields in the SAE/SUSAR CRF.
5. Generate SUSAR Fax.

SAE alert messages



ObTiMA_EURECASafe Study_New SAE has been reported

This is an automatically generated message sent by ObTiMA.
If you think you have received this message in error, please contact ObTiMA - Ontology-based Trial Management Application (administrator@obtima.org)!

Dear Mr. Random,

John Doe has recorded a new serious adverse event for patient *ECS_USAAR_001*.

For further details, log into ObTiMA and open the patients AE and SAE forms, respectively.

Please do not hesitate to contact ObTiMA - Ontology-based Trial Management Application if you need any assistance!

ObTiMA_EURECASafe Study_New SAE details available for ECS_USAAR_001

This is an automatically generated message sent by ObTiMA.
If you think you have received this message in error, please contact ObTiMA - Ontology-based Trial Management Application (administrator@obtima.org)!

Dear Mr. Random,

John Doe has entered new SAE details for *ECS_USAAR_001*.

For further details, log into ObTiMA and open the respective SAE form.

Please do not hesitate to contact ObTiMA - Ontology-based Trial Management Application if you need any assistance!

Figure 15: SAE Alerts

Log in

You have to log-in in the ObTiMA application first.

Open patient specific AE and SAE Form

The AE and SAE forms are deposited in the respective trial patient specific eCRF.

1. Select the clinical trial in the trials menu.
2. Select the patient, who is/has been affected by a serious adverse event.
3. Open the AE and SAE forms which are located in the respective study event (Figure 16).

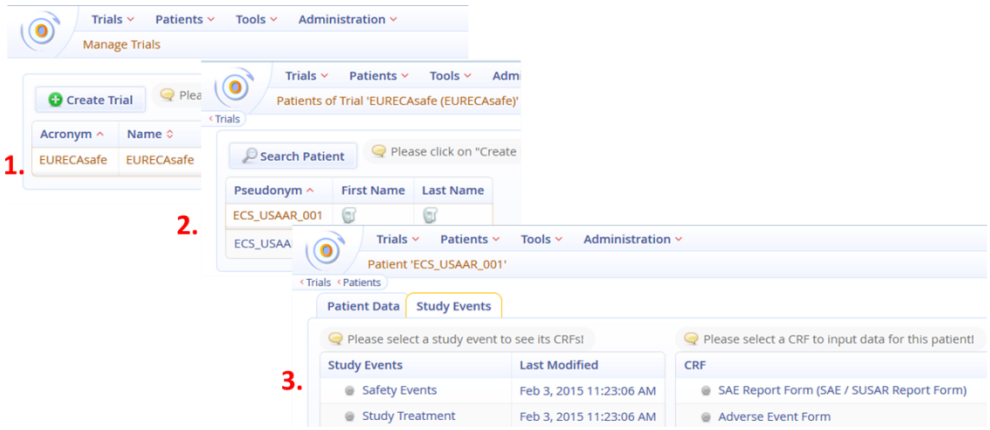


Figure 16: Patient specific AE/ SAE Form

Data validation of AE and SAE CRF

1. Check both forms for full data entry.
2. Perform source data verification (SDV) (Figure 17).
3. Carry out a medical review of the data (e.g. Query missing/ incorrect/ unclear items in the AE/ SAE CRFs).
4. After data has been verified and cleaned, lock the CRF data fields by clicking on the plus icon (⊕) next to the data field (see also 6.2)

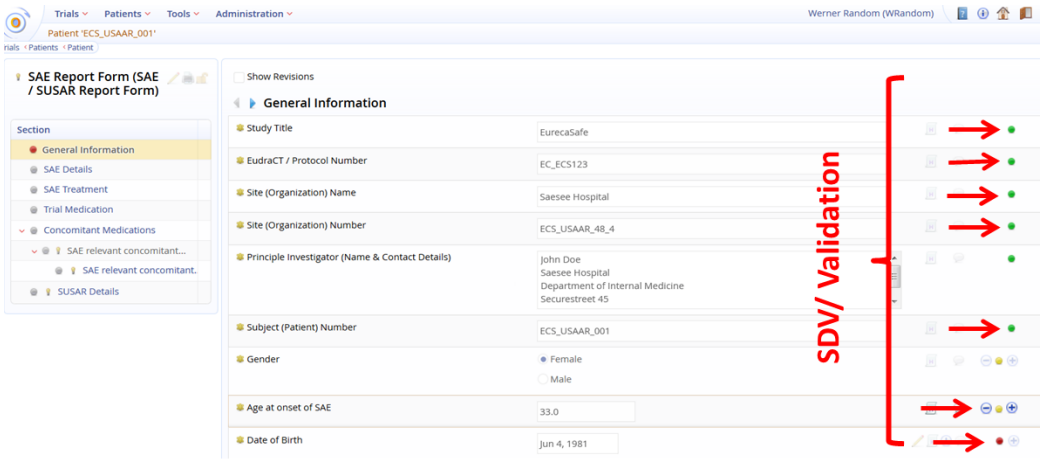


Figure 17: AE/ SAE Data validation

Assessment of causality and expectedness

The assessment of whether there is a reasonable possibility of a causal relationship between one or more IMPs (investigational medicinal product) and the event must always be assessed by local investigator and should not be downgraded by the sponsor/trial chairman. If the sponsor/ trial chairman disagrees with the investigator's causality assessment, the opinion of both the investigator and the sponsor should be provided using the comment function on the SAE CRF. The expectedness of an adverse reaction (AR) must be determined by the sponsor/ trial chairman on the basis of the reference safety information (Figure 18).

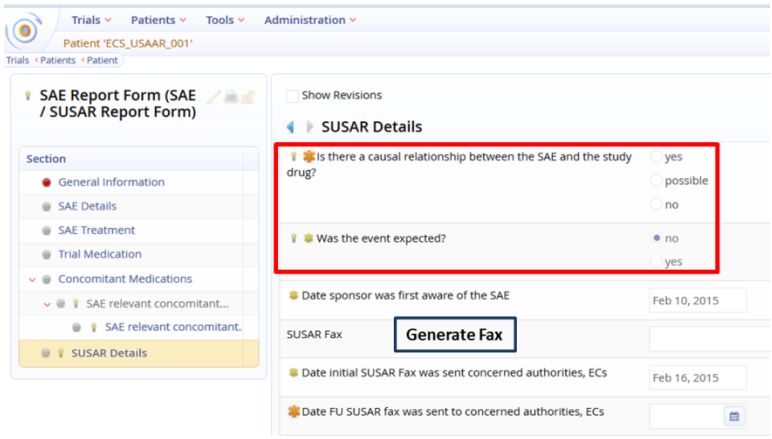


Figure 18: Evaluation of causality and expectedness

SAE/SUSAR Fax

1. After final confirmation, that the SAE is related and unexpected, click on the "Generate SUSAR FAX" button (Figure 19).
2. If all minimal required fields for the SAE/SUSAR Fax are available in ObTiMA, a pdf will be automatically generated. This pdf will be saved and can be opened afterwards (link to generated FAX).

NOTE: Mandatory data which is not been entered in the AE/ SAE CRF but elsewhere in ObTiMA (e.g. EudraCT number, sponsor study number, identifiable reporter) will be automatically inserted in the fax.

3. If not, a pop-up opens that informs about missing entries in ObTiMA. As far as all required fields are available, the fax can be generated (see 2).
4. Print the fax.
5. The document can be faxed to the EMA, the national competent authorities and to the Ethics Committees concerned. Please ensure to comply with the legally stipulated reporting timeline based on the SAE classification:
   - Fatal and Life threating in- between 7 days after knowledge
   - All others in-between 15 days
6. Enter the date of fax transmission in the field “SUSAR faxed to authorities/ Ethics committees” and save the CRF.

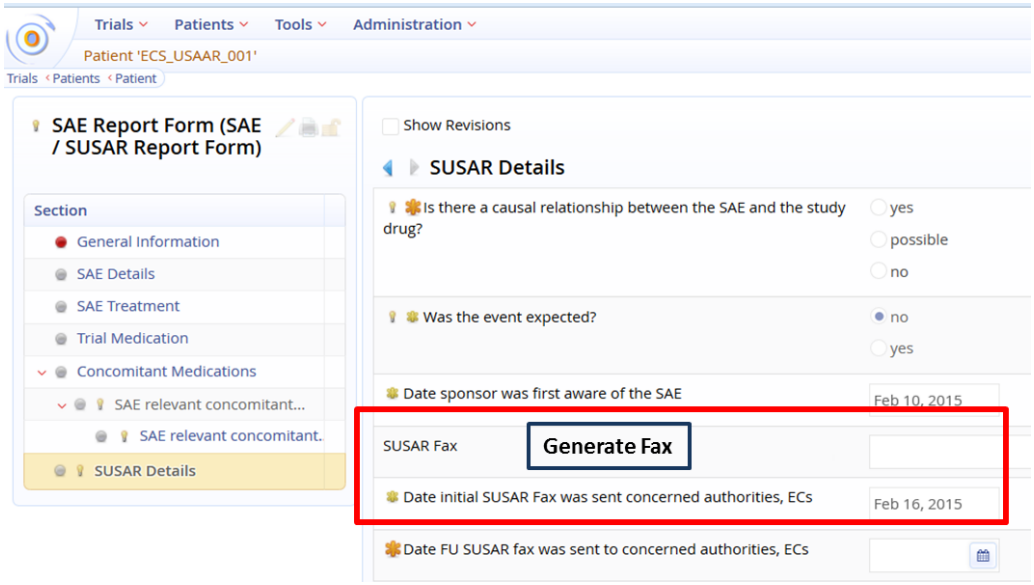


Figure 19: Generate SUSAR Fax

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 13: Main features of the evaluation procedure for the "Automatic detection and reporting of SAEs/SUSARs" scenario.

		Description
General	Evaluation Scenario	Automatic Detection of SAE
	Evaluation Leader	Norbert Graf
	Description of the scenario	Detect and report an episode of febrile neutropenia by extracting some specific symptoms and clinical relevant characteristics from EHR on a given period of time for retrospective study.

	Description of the evaluation procedure		Retrospective data from EHR are needed to detect SAEs and report them. It will be evaluated how exactly this will be possible.
Setup	Location of evaluation		UdS - Hospital for Paediatric Oncology and Haematology
	Setup details		
	Evaluator's expertise		Professor for Paediatric Oncology Haematology
	Number of evaluators		At least 3 physicians and 3 study nurses
	Duration of the evaluation		This will be done within 4 weeks after setup of the system.
	Time plan		April 2015 after the installation of the system
Data	Kind of data used for the evaluation		Retrospective data will be used.
	For retrospective data		Access internal data of the hospital. The evaluation is done in the clinical care situation.
	If you plan to access CDP servers		Not planned
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		The evaluation will be a proof of concept but also asking the question, if the tool can be used in clinical care.
Proof of Concept	What exactly will be evaluated?		It will be evaluated, if SAEs can be reported according to GCP criteria.
	How it will be evaluated?		The templates given in D8.1. will be used
	How many clinicians/users will be involved?		At least 3 clinicians and 3 study nurses
	How many times will they fill-in the usability questionnaires?		If the tool is not satisfying the end-users will run an iterative process with the developers to optimize the tool and a new evaluation will be done by the same end-users.
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation in an iterative process until the tool is satisfying all end-users. It will take place at the same location (USAAR) with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	A fictive trial will be set up in ObTiMA and fictive SAEs will be created to report SAEs and SUSARs. The detection of SAEs and SUSARs is not implemented in this tool.
		Inclusion/Exclusion Criteria?	None exclusion criteria. Included are all patients having an AE.

		Will the EURECA tool be compared with the tool you normally use?	No, as there is no corresponding tool available. Today every report of SAEs is done by hand.
		Which is the evaluation hypothesis?	The hypothesis is, that the tool will report SAEs according to GCP criteria.
	Evaluation factor 1	Quality metrics	No metrics
		Measured value	No values
		Rating levels	The rating given in D8.1 will be used.
	Criteria	Assessment criteria	Assessment criteria are the usability during clinical practice and the percentage of detected SAEs.

14 Microbiology SAE

The trial chairman defines in one or more specific CRFs which specific information has to be documented in order to get early knowledge about infectious agents and their resistance profile for usage of correct antibiotics in patients. Common Toxicity Criteria can be specified in order to detect SAE events automatically. These CRFs are summarized in a Microbiology Module (observational trial).

There are three main CRFs:

1. Clinical Data; entries like general values, patient's diagnoses, lab values, admission date, discharge date (linked to HIS)
2. Microbiology Data; entries spectrum of pathogens, infectious agents, antibiogram (linked to Microbiology database)
3. Antibiotic treatment

The Microbiology Module needs to be installed at the user site. This will be at UdS at the Hospital for Paediatric Oncology and Haematology.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 14: Main features of the evaluation procedure for the "Microbiology SAE" scenario.

		Description
General	Evaluation Scenario	Microbiology SAE
	Evaluation Leader	Norbert Graf
	Description of the scenario	The trial chairman defines in one or more specific CRFs which specific information has to be documented in order to get early knowledge about infectious agents and their resistance profile for usage of correct antibiotics in patients. Common Toxicity Criteria can be specified in order to detect SAE events automatically.
	Description of the evaluation procedure	The Microbiology Module needs to be installed at the user site. This will be at UdS at the Hospital for Paediatric Oncology and Haematology.
Setup	Location of evaluation	UdS - Hospital for Paediatric Oncology and Haematology
	Setup details	
	Evaluator's expertise	Professor for Paediatric Oncology Haematology
	Number of evaluators	At least 3 physicians and 3 study nurses
	Duration of the evaluation	This will be done within 4 weeks after setup of the system
	Time plan	Beginning of May 2015 after the installation of the system

Data	Kind of data used for the evaluation		Retrospective data will be used
	For retrospective data		Access internal data of the hospital. The evaluation is done in the clinical care situation.
	If you plan to access CDP servers		Not planned
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		The evaluation will be a proof of concept but also asking the question, if the tool can be used in clinical care.
Proof of Concept	What exactly will be evaluated?		It will be evaluated if the different components of the tool will work as expected, if the results provided by the tool are clinically relevant and helpful for the clinical care situation. In addition the tool will be tested according to the evaluation criteria as described in D8.1.
	How it will be evaluated?		The templates given in D8.1 will be used by the above mentioned 2 questions: Do the different components of the tool work smoothly together? Can the tool be used in daily clinical practice? Besides Yes and No to these two questions, free text will be given for more details.
	How many clinicians/users will be involved?		At least 3 clinicians and 3 study nurses
	How many times will they fill-in the usability questionnaires?		If the tool is not satisfying the end-users there will be an iterative process with the developers to optimize the tool and a new evaluation will be done by the same end-users.
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	Same end-users will do the evaluation in an iterative process until the tool is satisfying all end-users. It will take place at the same location (UdS) with the same parameters.
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	All patients that are on the ward of Paediatric Oncology at UdS that have microbiological analysis during the last week of the evaluation.

		Inclusion/Exclusion Criteria?	Inclusion: all patients with microbiological analyses. No exclusion criteria.
		Will the EURECA tool be compared with the tool you normally use?	There is no comparable tool available. The tool is innovative and needed by clinicians. No control group is available.
		Which is the evaluation hypothesis?	The hypothesis is, that the tool is helpful in early detection of resistant microbiological agents and therefore leads to a better and earlier correct antibiotic treatment.
	Evaluation factor 1	Quality metrics	As there is no comparable tool available one can only compare the time needed to get results from the tool, with the needed time to generate these results by hand, which is never stopped, as not done, but which is lasting for hours.
		Measured value	Time to get the results needed.
		Rating levels	Time below 10 minutes to get results is excellent.
	Evaluation factor 2	Quality metrics	
		Measured value	The question will be, if the tool is helpful in the clinical setting. The answer will be yes or no.
		Rating levels	According to the ranking as given in D8.1
	Criteria	Assessment criteria	Assessment criteria are the usability during clinical practice and the time needed to get results. If this is optimal, the next step will be to outsource the tool to all hospitals at USAAR and then to other hospitals in Germany.

15 Reporting episodes of febrile neutropenia

The scenario aims to detect an episode of febrile neutropenia in patients and to determine whether or not this episode is chemotherapy induced by automatically analysing patients' EHR data using adapted NLP tools.

One evaluation session will be done onsite in collaboration with the Statistics and Epidemiology Unit of IJB, with whom the tool is being developed. The system will be set-up locally at IJB.

The evaluators will ideally be composed of data managers, and one epidemiologist who has already fulfilled that task.

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 15: Main features of the evaluation procedure for the "Reporting episodes of febrile neutropenia" scenario.

		Description
General	Evaluation Scenario	Febrile neutropenia
	Evaluation Leader	IJB (Insitut Jules Bordet)
	Description of the scenario	The scenario aims to detect an episode of febrile neutropenia in patients and to determine whether or not this episode is chemotherapy induced by automatically analyzing patients' EHR data using adapted NLP tools.
	Description of the evaluation procedure	Evaluation session will be done onsite. The system will be set-up locally at IJB.
Setup	Location of evaluation	IJB (Insitut Jules Bordet)
	Setup details	Local set-up
	Evaluator's expertise	- Data managers - Epidemiologist
	Number of evaluators	- 1 data manager - 1 epidemiologist
	Duration of the evaluation	Around 1 hour per evaluation session
	Time plan	- Semi-structured interview - Result Performance
Data	Kind of data used for the evaluation	(1) Retrospective
	For retrospective data	(2) Access the CDP servers to perform the validation (the server of UPM, or of Custodix)

	If you plan to access CDP servers		(a) Yes (b) Not yet, but they will
	If you plan on using prospective data		Not planned
	Type(s) of evaluation		Proof of concept
Proof of Concept	What exactly will be evaluated?		We will mainly evaluate the information extraction results from free text data.
	How it will be evaluated?		We will compare automatic extraction with manual extraction, based on performance index (e.g. sensitivity/recall, specificity, precision measurements and calculation), and measure the time required with both methods (manual and automatic).
	How many clinicians/users will be involved?		2 or 3
	How many times will they fill-in the usability questionnaires?		The development process will be iterative, but we think we will only have time for one formal evaluation session, as the tools developed in this scenario can be seen as an extension of the ones developed for the scenario on "Cancer registry reporting".
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	(3) Retrospective
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	The tool will be tested on at least 20 patients from the Breast free text dataset.
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	The tool will be compared with usual manual extraction.
		Which is the evaluation hypothesis?	The evaluation hypothesis is to compare automatic and manual extraction, and assess the time performances.
	on	Quality metrics	Time

		Measured value	Time required to extract and fill-in relevant information in the cancer registry.
		Rating levels	The difference between manual extraction time and automatic extraction time will be measured.
	Evaluation factor 2	Quality metrics	Sensitivity/recall, specificity, precision, ROC curves.
		Measured value	The quality metrics will be measured on all values of relevant concepts that are meant to be extracted with the tool. The correct/incorrect extracted values will be judged by the users.
		Rating levels	
	Criteria	Assessment criteria	Sensitivity/recall, specificity, precision, ROC curves.

16 Cancer registry and tumour bank reporting

The scenario aims to fill part of the internal cancer registry by automatically extracting information related to the categorization of any new tumour case (e.g. the morphology, the topography, the tumour staging) out of patients' EHR data using adapted NLP tools. Set-up in 2000, the objective of the registry is to record structured information about all cancers diagnosed and/or treated at the Institute (presently more than 33,000 cases). The objectives of the project are multiple: describing patient profiles and tumour characteristics; assessing the Institute's medical activity and patient prognosis; and contributing to research and the national Cancer Registry. Procedures, specifically those for quality assurance and case identification, are developed continually.

Several evaluation sessions will be done iteratively onsite in collaboration with the Statistics and Epidemiology Unit of IJB, with whom the tool is being developed. The system will be set-up locally at IJB.

The evaluators will ideally be composed of data managers, one epidemiologist and one statistician who are already working on filling or exploiting entries in the cancer registry. The final evaluation sessions are expected to be hold between May and July 2015 at IJB.

The evaluation procedure will consist usage of the tool in order to fill automatically specific part of the cancer registry, and to check manually whether such filled information is correct or not will looking at the EHR information sources, in both dataset that have been provided by IJB within the EURECA project (Breast free text dataset, and Breast structured dataset).

The following table reports the main features of the evaluation procedure for the specific scenario.

Table 16: Main features of the evaluation procedure for the "Cancer registry and tumour bank reporting" scenario.

		Description
General	Evaluation Scenario	Cancer registry reporting
	Evaluation Leader	IJB (Institut Jules Bordet)
	Description of the scenario	Set-up in 2000, the objective of the registry is to record structured information about all cancers diagnosed and/or treated at the Institute (presently more than 33,000 cases).

	Description of the evaluation procedure	Several evaluation sessions will be done iteratively onsite in collaboration with the Statistics and Epidemiology Unit of IJB, with whom the tool is being developed. The system will be set-up locally at IJB. The evaluators will ideally be composed of data managers, one epidemiologist and one statistician who are already working on filling or exploiting entries in the cancer registry.
Setup	Location of evaluation	IJB (Institut Jules Bordet)
	Setup details	Local set-up
	Evaluator's expertise	- Data managers - Epidemiologist - Statistician
	Number of evaluators	- 1 or 2 data managers - 1 epidemiologist - 1 or 2 statisticians
	Duration of the evaluation	Several evaluation sessions will be done iteratively onsite in collaboration with the Statistics and Epidemiology Unit of the Institut Jules Bordet, with whom the tools is developed.
	Time plan	- Semi-structured interview - Result Performance
Data	Kind of data used for the evaluation	(1) Retrospective
	For retrospective data	(2) Access the CDP servers to perform the validation (the server of UPM, or of Custodix)
	If you plan to access CDP servers	(a) Yes (b) Not yet, but they will
	If you plan on using prospective data	Not planned
	Type(s) of evaluation	Proof of concept
Proof of Concept	What exactly will be evaluated?	We will mainly evaluate the information extraction results from free text data.
	How it will be evaluated?	We will compare automatic extraction with manual extraction, based on performance index (e.g. sensitivity/recall, specificity, precision measurements and calculation), and measure the time required with both methods (manual and automatic).

	How many clinicians/users will be involved?		Between 3 and 5
	How many times will they fill-in the usability questionnaires?		The development process will be iterative, so we plan to make the evaluation users fill-in the usability questionnaire several times (at least twice).
Clinical Evaluation	**Set-Up**	Clinical Evaluation Type	(3) Retrospective
		How many patients/clinicians will be tested? How will they be selected? Are they going to sign an informed consent?	The tool will be tested on at least 20 patients from the Breast free text dataset.
		Inclusion/Exclusion Criteria?	None
		Will the EURECA tool be compared with the tool you normally use?	The tool will be compared with usual manual extraction.
		Which is the evaluation hypothesis?	The evaluation hypothesis is to compare automatic and manual extraction, and assess the time performances.
	Evaluation factor 1	Quality metrics	Time
		Measured value	Time required to extract and fill-in relevant information in the cancer registry.
		Rating levels	The difference between manual extraction time and automatic extraction time will be measured.
	Evaluation factor 2	Quality metrics	Sensitivity/recall, specificity, precision, ROC curves.
		Measured value	The quality metrics will be measured on all values of relevant concepts that are meant to be extracted with the tool. The correct/incorrect extracted values will be judged by the users.
		Rating levels	
	Criteria	Assessment criteria	e.g. sensitivity/recall, specificity, precision, ROC curves.

17 CONCLUSIONS

In this document we report the evaluation procedures to be followed in the clinical sites for the evolution of the 15 clinical scenarios. The clinical sites in EURECA are five (5). UOXF leads five (5) clinical scenarios for evaluation, UdS leads five (5) clinical scenarios, IJB and Maastro lead two (2) clinical scenarios and GBG lead one (1) clinical scenario as shown in Figure 20.

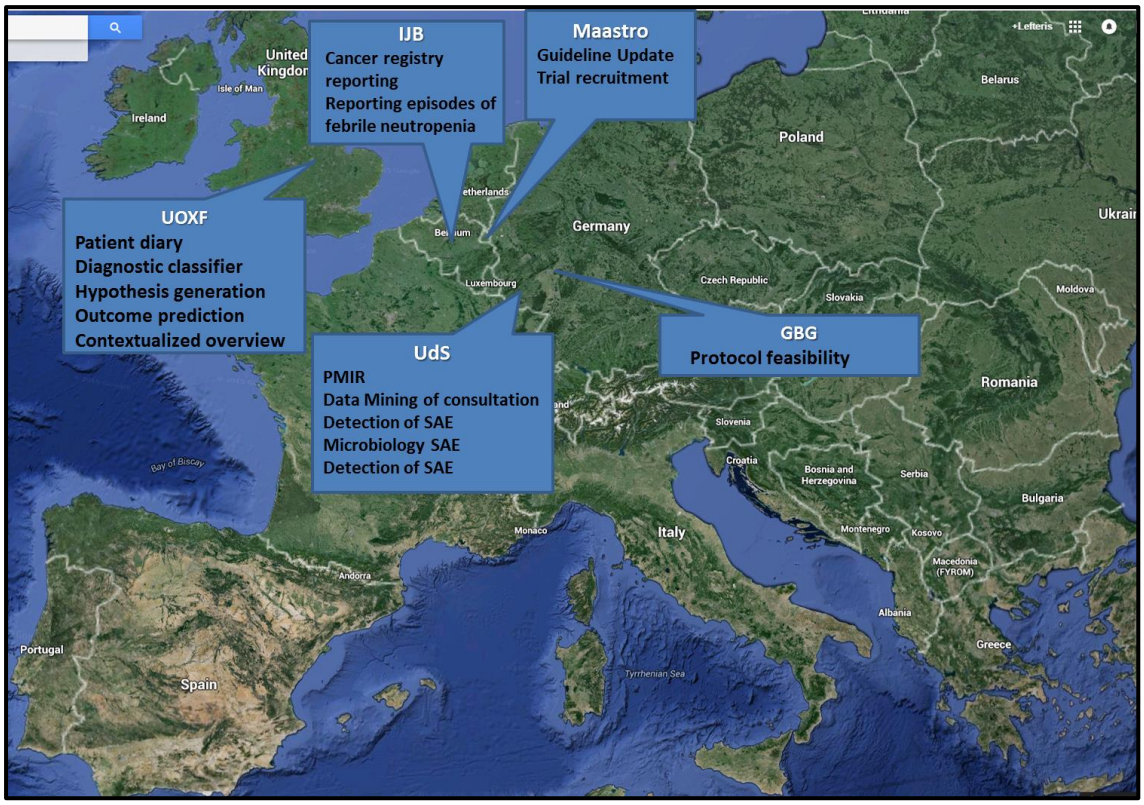


Figure 20: Clinical sites and the scenarios they lead

For the effective execution of the worked planned, we will also focus on delivering the required training to end users, through structured user workshops to be held at each clinical site. At least one workshop in each clinical site is foreseen in order to train the end users and assure a smooth and effective evaluation. The results of the training will be reported in Deliverable D8.6.

In order to achieve a qualitative evaluation for all the scenarios, we will recruit experts from many fields. Different end users are required for different use cases and scenarios. Table 17 summarizes the number and the expertise of the evaluators in each clinical scenario.

Table 17: Evaluators per scenario

Leader	Scenario	Clinicians	Research nurses	Biostatisticians/ Bioinformaticians	Other
UdS	Personal medical information recommender	3	3		

	Data mining of consultation	1			
	Prediction of SAEs/SUSARs	3	3		
	Automatic detection and reporting of SAEs/SUSARs	3	3		
	Microbiology SAE	3	3		
UOXF	Contextualized overview	3 or 4			
	Patient Diary & Long-term follow-up	3 or 4			
	Hypothesis generation			3	
	Outcome prediction			3	
	Diagnostic classifier			3	
IJB	Reporting episodes of febrile neutropenia				1 Epidemiologist & 1 Data manager
	Cancer registry and tumor bank reporting			1 or 2	1 Epidemiologist & 1-2 Data managers
Maastro	Update of guidelines	1			1 Researcher & 1 Guideline designer
	Trial recruitment		2		
GBG	Protocol feasibility	1			1 Study manager

Another crucial aspect in the evaluation of clinical infrastructure is the availability and the quality of data. EURECA semantic layer and EURECA Data Protection Framework assure the quality and the availability of the data even in cases where strict restrictions apply. Table 18 summarizes the type of data and the type of access. As shown, only two of the clinical scenarios will use prospective data as the collection of this type of data is practically impossible for the most of the clinical scenarios due to the lack of time.

Table 18: Data to be used for the evaluation

Leader	Scenario	Retrospective	Prospective
UdS	Personal medical information recommender	Fake data of patients entered in the EHR (Indivohealth™)	
	Data mining of consultation	Retrospective and anonymized data of the consultation CRF of SIOP 2001	
	Prediction of SAEs/SUSARs	SIOP 2001 data, if the prediction will be done for VOD, accessed through the CDP servers.	
	Automatic detection and reporting of SAEs/SUSARs	Access internal data of the hospital. The evaluation is done in the clinical care situation.	
	Microbiology SAE	Access internal data of the hospital. The evaluation is done in the clinical care situation.	
UOXF	Contextualized overview	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).	

	Patient Diary & Long-term follow-up	Access the CDP servers to perform the validation (the server of UPM, or of Custodix)	
	Hypothesis generation	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).	
	Outcome prediction	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).	
	Diagnostic classifier	Access the CDP servers to perform the validation (the server of UPM, or of Custodix).	
IJB	Reporting episodes of febrile neutropenia	Access the CDP servers to perform the validation (the server of UPM, or of Custodix)	
	Cancer registry and tumor bank reporting	Access the CDP servers to perform the validation (the server of UPM, or of Custodix)	
Maastro	Update of guidelines		✓
	Trial recruitment	Patient data from MAASTRO and German Breast Group	
GBG	Protocol feasibility		✓

Of course the evaluation is based on specific criteria. We have defined all the evaluation criteria for each scenario which are also summarized in Table 19.
Criteria include but are not limited to:

- Comparison (e.g. prediction with the recommendation of a clinician)
- Statistics (e.g. Sensitivity/recall, specificity, precision, ROC curves)
- Usability evaluation (using SUS questionnaire and Evaluation Questionnaire based on the 25000 ISO series)
- Time (e.g. time it takes to do a specific job compared to current procedures)
- Minimum acceptance levels

Table 19: Evaluation criteria

Leader	Scenario	Evaluation criteria
UdS	Personal medical information recommender	Comparing the prediction with the recommendation of a clinician
	Data mining of consultation	(1) Average understandability of topics, the number of understandable and the number of understandable or somewhat understandable topics, and the number of relevant or somewhat relevant topics amongst all understandable topics. (2) Number of frequent or somewhat frequent questions in the list, plus the number of too general or too specific candidates. (3) The fraction of recommended FAQs that are indeed relevant, and the number of recommended similar queries that are indeed similar.

	Prediction of SAEs/SUSARs	Usability of the tool and accuracy of the prediction models
	Automatic detection and reporting of SAEs/SUSARs	Usability during clinical practice and the percentage of detected SAEs
	Microbiology SAE	(1) Compare the time needed to get results from the tool, with the needed time to generate these results by hand (2) If the tool is helpful in the clinical setting. The answer will be yes or no.
UOXF	Contextualized overview	(1) For filtering use cases, the application is either incapable of returning a result, or it returns a correct result. As such, the number of cases in which it the application is capable of returning an answer is measured. (2) For the ranking use cases, a comparison in time will be made between the manual formulation of a query, and the querying process the system will execute. Qualitative assessment will be described below (3) For the ranking use cases, there is no golden standard to compare the generated results to. Therefore a more subjective comparison will be made by measuring the number of titles on which a user clicks up to a certain number, after both a manual formulation of a query and the query process executed by the system.
	Patient Diary & Long-term follow-up	(1) Time to create a new event, or a new appointment on the calendar (2) Time of inserting, editing deleting a procedure and a medication (3) Accuracy of the drug-drug interaction
	Hypothesis generation	(1) Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc. (2) The number of alternative paths that analysts are able to explore, i.e. with the help of the tool how many possible ways can be thought of to generate different hypotheses by the same subject. (3) The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task
	Outcome prediction	(1) Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc. (2) The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task.

		(3) Performance (3.1) For prediction of binary outcome: AUC = Area under the ROC (receiver operating characteristic) curve (3.2) For prediction of survival outcome: c-index = concordance index, a performance measure specifically for censored data
	Diagnostic classifier	(1) Time it takes to comprehend the data mining task and translate it to the actions in KDF. E.g. given an objective, how long does it take to select the dataset, how long to modify the script, how long to run the script, etc. (2) The number of concepts that can be considered in the analysis, i.e. with the help of the KDF how many clinical concepts from CDM are useful for the data mining task. (3) Robustness of clustering: a similarity measure for the patient clusters if cross-validation or bootstrapping procedures are applied
IJB	Reporting episodes of febrile neutropenia	(1) Time required to extract and fill-in relevant information in the cancer registry. (2) Sensitivity/recall, specificity, precision, ROC curves
	Cancer registry and tumor bank reporting	(1) Time required to extract and fill-in relevant information in the cancer registry. (2) Sensitivity/recall, specificity, precision, ROC curves
Maastro	Update of guidelines	(1) How many articles are relevant to the guideline update. (2) Percentage of discovered articles deemed to be relevant.
	Trial recruitment	Usability score from the usability questionnaire
GBG	Protocol feasibility	Correctness, Timing, Usability, the outcome of the questionnaires A and B

Evaluations will be conducted between month 38 and month 40 of the project. An overview of the scheduled timeline for the evaluation of all the scenarios is shown in Table 20. Furthermore, in the table we can identify which evaluations are proofs of concept (marked with blue), which are clinical test (marked with green) and which are algorithmic performance evaluations (marked with orange). Clinical tests and algorithmic performance evaluations will also support usability evaluations using questionnaires (SUS and ISO 25000 series based questionnaire). As we can see many scenarios will be evaluated as clinical test with procedures similar to real clinical practice. Such evaluations span beyond the usability and provide us the option to demonstrate the real impact of the EURECA infrastructure.

Table 20: Type of evaluation (blue indicates Proof of concept, green Clinical Test)

Leader	Scenario	April (m39)				May (m40)			
UdS	Personal medical information recommender								
	Data mining of consultation								
	Prediction of SAEs/SUSARs								
	Automatic detection and reporting of SAEs/SUSARs								
	Microbiology SAE								
UOXF	Contextualized overview								
	Patient Diary & Long-term follow-up								
	Hypothesis generation								
	Outcome prediction								
	Diagnostic classifier								
IJB	Reporting episodes of febrile neutropenia								
	Cancer registry and tumor bank reporting								
Maastro	Update of guidelines								
	Trial recruitment								
GBG	Protocol feasibility								

Type of evaluation

Proof of concept

Clinical Test & proof of concept

Algorithmic Performance evaluation & proof of concept

Results of the evaluation will be reported in the deliverable D8.5 (Report on the evaluation and validation of the EURECA environment and services) provided to the developers for further refinement and updates.

18 REFERENCES

[1] EURECA consortium, D1.2 Definition of relevant user scenarios based on input from users, 2013
[2] EURECA consortium, D1.1 User needs and specifications for the EURECA environment and software services, 2012
[3] Scholtz, Jean. "Beyond usability: Evaluation aspects of visual analytic environments." Visual Analytics Science And Technology, 2006 IEEE Symposium On. IEEE, 2006
[4] DIRECTIVE 2001/20/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use (OJ L 121, 1.5.2001, p. 34)
[5] Detailed guidance on the collection, verification and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use ('CT-3') (2011/C 172/01)
[6] ICH HARMONISED TRIPARTITE GUIDELINE MAINTENANCE OF THE ICH GUIDELINE ON CLINICAL SAFETY DATA MANAGEMENT: DATA ELEMENTS FOR TRANSMISSION OF INDIVIDUAL CASE SAFETY REPORTS E2B(R2) Current Step 4 version dated 5 February 2001

Appendix A

		Description	Guidelines
General	Evaluation Scenario		
	Evaluation Leader		
	Description of the scenario		**Detailed description** of the scenario (will be part of the scenario's section as introduction)
	Description of the evaluation procedure		**Detailed description** of the evaluation procedure (will be part of the scenario's section as introduction and will be removed from the template)
Setup	Location of evaluation		
	Setup details		e.g. set-up locally, remote access etc.
	Evaluator's expertise		e.g. patient, physician, nurses
	Number of evaluators		
	Duration of the evaluation		Time needed after the set-up of the system
	Time plan		
Data	Kind of data used for the evaluation		Possible choices (1) Retrospective (2) Prospective (3) Both (4) None (if none please explain)
	For retrospective data		Possible Choices (1) Access the internal data of your institution (2) Access the CDP servers to perform the validation (the server of UPM, or of Custodix) (3) Other – If so, please explain
	If you plan to access CDP servers		**All** following bullets should be answered (a)Are all the people accessing the data part of your organization ? (b) Did all of them sign Annex C ?
	If you plan on using prospective data		**All** following bullets should be answered (a) Did the validation undergo the review process of an ethics board, or any other approval procedure as expected by your national law (b) Did you, or will you obtain consent for

			processing of data from the patients?
	Type(s) of evaluation		clinical testing OR proof of concept. **In case of proof of concept the evaluator must prepare and deliver evaluation and usability questionnaires** and ignore the evaluation factors in this template
Proof of Concept	What exactly will be evaluated?		Please describe what you will evaluate in details
	How it will be evaluated?		Please describe how you will perform the aforementioned evaluation
	How many clinicians/users will be involved?		
	How many times will they fill-in the usability questionnaires?		Is there going to be an iterative development process?
Clinical Evaluation	**Set-Up** Clinical Evaluation Type		(1) Sequential: Same end users, same location(s), same parameter(s), Duration sequential (X months without EURECA system and Y months with EURECA system. (2) Parallel: Duration parallel, with and without the EURECA environment, Random distribution of cases/samples to the end users (3) Retrospective: Reproduce the study using EURECA environment and compare
	Set-Up How many patients/clinicians will be tested? How will they be selected? Are they going to sign		

		an informed consent?		
		Inclusion/Exclusion Criteria?		
		Will the EURECA tool be compared with the tool you normally use?		Is there going to be a control group? Which group is that?
		Which is the evaluation hypothesis?		
	Evaluation factor 1	Quality metrics		Metrics must correlate to the characteristics of the EURECA scenario. Every quantifiable feature of software and every quantifiable interaction of software with its environment that correlates with a characteristic can be established as a metric. Metrics can differ depending on the environment and the different end user groups.
		Measured value		Producing values that are (1) formal enough to serve as a basis for comparison amongst alternative methods under consideration; (2) mappable to utility e.g., measuring the weight of some object of evaluation should only happen if it is clear how weight relates to utility
		Rating levels		Rating level, intervals, minimum accepted threshold have to be determined
	…	…..		
	Evaluation factor N	Quality metrics		

	Measured value		
	Rating levels		
Criteria	Assessment criteria		The evaluation leader has to prepare a procedure for this, using, for instance, decision tables, rules or weighted averages. The procedure usually will include other aspects such as time and cost that contribute to the assessment of quality of the scenario in the particular environment.