

## **Work Plan**

**Allard van Altena, V4**

E-mail: [allard@van-altena.net](mailto:allard@van-altena.net)

Phone: 06-3 1683756

### **Title**

Developing and Deploying a Data Management System for Obstetric Research into (long) term effect of In Vitro Fertilisation techniques

### **Place of the SRP Project**

Department of Clinical Epidemiology, Biostatistics and Bioinformatics

Department of Obstetrics and Gynaecology

Department of Medical Informatics

### **Mentor**

Dr. Silvia Delgado Olabarriaga, AMC, Department of Clinical Epidemiology, Biostatistics and Bioinformatics

[s.d.olabarriaga@amc.uva.nl](mailto:s.d.olabarriaga@amc.uva.nl)

### **Tutor**

Dr. Ir. A.C.J. Ravelli, AMC, Department of Medical Informatics

[a.c.ravelli@amc.uva.nl](mailto:a.c.ravelli@amc.uva.nl)

### **Contact persons**

Clinical support: Prof. Dr. T.J. Roseboom, AMC, Department of Obstetrics and Gynaecology

[t.j.roseboom@amc.uva.nl](mailto:t.j.roseboom@amc.uva.nl)

### **Description of the SRP Project/Problem**

Until now no large scale studies have been conducted which differentiate between different IVF procedures on neonatal outcomes. Therefore it is unknown what the effects of a specific procedure are on a new-born (and later, adult). These studies will only become possible when combining data from IVF clinics with various other data sources, one of which is the PRN. The department of obstetrics and gynaecology is starting to collect, organise, and link these types of data. However, various challenges need to be faced before a study can be conducted.

The data separation described in the introduction is not the main problem addressed in this study. The method for data linkage will be provided by another project. At this point, when a data linking method is available, we run into the following problems: data heterogeneity, data scattering, data collection, data security, data access, data browsing, data querying. From this we can deduce that we have a problem related to the field of "big data". The current workflow for this dataset is to link it manually and test one or several hypothesis on the linked data. Afterwards the dataset is mostly discarded and not used for further research. What we want to achieve is to increase the efficiency for the researcher when using this dataset. Increasing efficiency can be achieved by providing tools for the researcher to access and use the dataset. This research will therefore be about how to successfully deploy a system implementing tools solving the previously mentioned problems.

### **Research Questions**

1. How do we implement a user-friendly system which covers problems concerning: data heterogeneity, data scattering, data collection, data security, data access, data browsing, and data querying?
  - a. What are the functions of this system, e.g.: automatic and dynamic collection and linkage of data, view data, provide data to end-user eligible for use in research, view data usage history, provide statistical testing as a service (SPSS server), provide secure access to the data?
  - b. What are the legal aspect of the data collection, storage and analysis (e.g. traceability to the individual, integrity, IVF centre specific conditions, PRN specific conditions, and what has been agreed about the linked data)?
  - c. What are the main criteria for a data model which can differentiate between IVF procedures considering new-borns?
  - d. What is the data model for PRN data, IVF clinic data, and additional data sources?
  - e. What is the data model for the linked dataset?
  - f. Who are the users and how should these data models be presented to these users?
  - g. How can data mining be applied to the system and deliver valuable functionality for the end-

user?

2. To what extent does the system help to answer the following clinical question: Does the type of culture medium used make a difference on the course and outcome of pregnancy?

### Planning of the Project Activities

1. Mind storm (with key stakeholders, e.g.: end-users of the system) about possible functionality for the system and make a map out of all the ideas, e.g.: data visualisation, data provenance/history, providing data analysis (SPSS) as a service, providing data mining as a service. From this set select feasible ideas for the execution of this SRP.
  - **Method:** Either a mind storm method will be used or a regular discussion where certain pre-defined questions are stated by me (A. van Altena), results will be written down by hand.
  - **Deliverable:** Write these ideas down in a short mind map to refer to later on in the project.
  - **Deadline:** 7<sup>th</sup> of February.
  - **Answers question:** 1a, 1f.
2. Create a focus group of experts who will be tied closely to the execution of the development and testing phase, could include stakeholders from step 1.
  - **Deadline:** 7<sup>th</sup> of February.
  - **Dependent on:** 1.
3. Data collection, working together with the assigned PhD student to gather data from the clinics. This will be done in person. There are a total of thirteen clinics a number of these will not deliver a set of data, for these clinics a data query will have to be created.
  - **Deadline:** 28<sup>th</sup> of February.
  - **Deliverable:** Dataset containing all data for clinics where it was possible to gather data. Also, a description of data collection process and encountered problems and how these were overcome.
4. Determine which types of data are going to be used in this SRP in order to fine-tune the level of security needed specifically to this use case.
  - **Method:** Literature study and interviews.
  - **Deliverable:** Data model/scheme used for data storage. A short report will be written as a referral for the rest of the SRP.
  - **Deadline:** 7<sup>th</sup> of February.
  - **Answers question:** 1c, 1d, 1e.
5. Make a study of the security issues which are present when developing a system around IVF and PRN data, e.g.: legal issues, IVF centre specific conditions, and PRN specific conditions.
  - **Method:** Literature study and interviews.
  - **Deliverable:** Documentation of these issues as a base for later requirement analysis and checklist for system prototype for testing purposes.
  - **Deadline:** 28<sup>th</sup> of February.
  - **Dependent on:** 4.
  - **Answers question:** 1b.
6. Study existing security solutions for the found legal and/or security issues and document them. For example: what security is used with the PRN data?
  - **Method:** Literature study and interviews.
  - **Deliverable:** Documentation of proven solutions from literature and real-life examples.
  - **Deadline:** 31<sup>st</sup> of March.
  - **Dependent on:** 4, 5.
  - **Answers question:** 1b.
7. *The steps 8, 9, 10, and 11 are performed in an iterative loop until the goal of a properly functioning prototype is reached.*
  - **Method:** Agile development method.
  - **(Overall) Deliverable:** Code and process documentation.
  - **Deadline:** *First iteration:* 31<sup>st</sup> of March, *Last iteration:* 31<sup>st</sup> of May.
  - **Answers question:** *fine tuning* 1a, 1c, 1d, 1e, and 1f.
  - **Dependent on:** 1, 2, 4, 5, 6.
8. Analyse requirements concerning the data used in the system. Describe the data used and what processes are needed for a functioning system. For example processes for data inflow (from the clinics and PRN) and data outflow (to researchers and back to clinics) are needed. Also find and describe system requirements and processes which are not directly related to the data.

- **Method:** Software engineering requirement analysis, mind storm with focus group/key stakeholders.
  - **Deliverable:** Requirement analysis report, UML description of system.
  - **Dependent on:** 1, 2, 4, 6.
9. Determine key requirements and key processes by applying the MoSCoW-method to the list of requirements.
- **Method:** MoSCoW-method.
  - **Deliverable:** Annotated requirement analysis report with MoSCoW analysis.
  - **Dependent on:** 1, 2, 4, 6, 8.
10. Develop and test key system processes, hereby trying to pull apart the complexity of the system into separate modules which makes testing easier and reduces risk of security issues. Also develop and test back-end and front-end (user-interfaces) incorporating key system processes (abstract algorithms). Pulling together the system modules handling the key system processes into a coherent user-friendly system through user-driven design approach.
- **Method:** Agile developing methods (to be determined).
  - **Deliverable:** System modules in the most suitable programming language accompanied with unit tests.
  - **Dependent on:** 1, 2, 4, 6, 8, 9.
11. Evaluate the system with unit tests and end-users. Based on the outcomes of this evaluation it is decided whether or not to perform another iteration.
- **Method:** User-driven design and unit tests.
  - **Dependent on:** 1, 2, 4, 6, 8, 9, 10.
12. Test the system with a real life research case.
- **Deadline:** 30<sup>th</sup> of June.
  - **Method:** Statistical analysis (to be determined, e.g. type of analysis is based on type of data).
  - **Deliverable:** Outcome data usable for writing scientific paper answering the SRP research question number 2, working system with an optimised user-interface.
  - **Dependent on:** 1, 2, 4, 6, 7, whole of 7.
13. With the outcomes of the case from step 12 write a scientific paper to comply with the Medical Informatics master module 10.
- **Deliverable:** Scientific paper concerning SRP research question number 2.
  - **Deadline:** 30<sup>th</sup> of June.
  - **Dependent on:** 1, 2, 4, 6, 7, whole of 7.
  - **Answers question:** 2.
14. Determine data mining method which can be applied to the created data model.
- **Method:** Research into existing data mining methods.
  - **Deliverable:** Definition of data mining method possibly in the form of a pseudo code algorithm.
  - **Deadline:** 30<sup>th</sup> of June.
  - **Answers question:** 1g.
  - **Dependent on:** 1, 2, 4.
15. Write master thesis encapsulating all deliverables describing the process which was used to get from the start of the SRP to the end.
- **Deliverable:** Master thesis.
  - **Deadline:** 31<sup>st</sup> of July.
  - **Dependent on:** All.
  - **Answers question:** All.
16. Release system as a prototype for use by end-users.
- **Deliverable:** Functioning prototype of the system with user documentation (end-user and system administrator).
  - **Deadline:** 31<sup>st</sup> of July.
  - **Dependent on:** 1, 2, 4, 6, whole of 7.

Table

MONTH	WEEK 1	WEEK 2	WEEK 3	WEEK 4
January				
February	1: Mindstorm, 2: Focus group, 4: Data types			3: Data collection, 5: Security issue study
March				6: Security solutions, 7: First iteration
April				
May				7: Last iteration
June				12: Case test, 13: Case test paper, 14: Data mining methods
July				15: Master thesis, 16: Release of system

#### Time period

48 ECTS, i.e. 32 weeks of 36 working hours  
1 December 2014 – approximately 31 July 2014


#### Special Circumstances

The execution of this SRP will have some risks. One of the biggest risks is the case where not enough data can be retrieved from the clinics to work with. At the beginning of the project it was not clear that there was not enough data to carry the SRP. For this case dummy data will have to be used, this can be created by using random output from a piece of code. However, research question number 2 needs actual data, this question can be adapted for the available data.

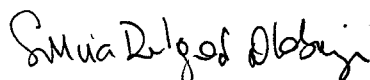
Other risks have mostly to do with the availability of people needed for one of the steps. For example, step 3 includes the PhD student for data collection, or step 2 which includes multiple people. If in one of these cases the required people cannot be reached I will try to find a solution on the go.

Lastly there are two technical risks. The first is that the security of the system cannot be guaranteed under the strict rules that apply at the AMC. The other risk is that no suitable (or affordable) equipment can be found for running the system. As this will be a prototype system it does not necessarily need to run on a server, or run with the actual data (dummy data can be used for presentation purposes). For the purpose of step 12 the system can be run locally (on a desktop/laptop) with the actual data grabbed from an encrypted USB-stick, in that case the data will never be stored in an unsecured system and it will never be send through a network.

#### Student



#### Mentor



#### Tutor

