# CLOUDSTUDY: A Cloud-Based System for Supporting Multi-Centre Studies

Amalia Tsafara, Christos Tryfonopoulos and Spiros Skiadopoulos

*Abstract*— **Among the basic research tools for (bio)medical science are epidemiological studies that typically involve a number of hospitals, clinics, and research centres scattered around the world, and are often referred to as multi-centre studies. Clearly, the effectiveness and importance of a multi-centre study increases with the number of participating centres and enrolled patients, but at the same time this natural distribution in the production of research data requires sophisticated data management infrastructures to support the participating units. This kind of infrastructure is not only expensive to build and maintain, but also cannot be reused as it is often tailored to a specific study. In this work, we present a cloud-based system, coined CLOUDSTUDY, that allows users without any computer science background to design, deploy, and administer platforms aimed for managing, sharing, and analysing clinical data from multi-centre studies. The CLOUDSTUDY system provides a zero-administration, zero-cost online tool for creating multi-centre studies that (i) enhances re-usability by introducing study templates, (ii) supports (bio)medical needs through specialised data types, and (iii) emphasises data filtering/export through an expressive yet simple graphical query engine.**

## I. INTRODUCTION

Among the basic research tools for (bio)medical science are large-scale multi-centre studies, that often involve a number of different stakeholders including hospitals, clinics, and research centres. These multi-centre studies are useful for drawing conclusions on a number of important research questions, but at the same time pose a number of issues including the collection, organisation, and processing of data (that are naturally *distributed* and produced *asynchronously*). To tackle with data fragmentation and solve the issues arising from the coordination of geographically distributed participants, a number of platforms (e.g., [12], [4], [11], [14], [5], [6], [13], [10]), that focus on the storage and management of (bio)medical data, have been proposed. However, all these platforms are either designed for a *specific task or study* [12], [6], [13], [8], [1], [3] (and are thus unusable in any other study), or require an *expert* in Information Technology (IT) and significant *computing infrastructure* for setup and tuning [10]. This results in (i) time-consuming meetings between scientists of different principles trying to understand each other's needs and (ii) resource-consuming IT infrastructure, that requires outsourcing to IT specialists and regular maintenance/upgrades to keep up with technological requirements. Due to these issues, a great number of multi-centre studies that lack the resources are still performed by resorting to *manual procedures*, such as collecting data on paper, exchanging data by post, or emailing enormous spreadsheet files with patient data. Therefore, concerns like coordination among participants, data freshness/integrity, control of participants' involvement, and timeliness of results are lost between versioning in exchanged spreadsheets, hard copies of patient data, and requests for participation.

**Idea and Challenges.** In this work, we present CLOUDSTUDY a *cloud-based* service for *managing, sharing, and organising* clinical and patient data from *multi-centre studies*. The proposed system covers all the functional requirements posed by multi-centre studies, and enables researchers to easily organise and share data and knowledge generated by the research activity. We propose an innovative integrated framework for creating platforms for multi-centre studies that enables users with *no prior IT knowledge* to (i) *design and launch*, in an easy and transparent way, platforms tailored to the specific needs of their studies, (ii) perform basic and advanced *user management* tasks (manage users, assign user privileges and permissions, perform access control on data), (iii) *record, organise and manage* clinical/patient data by resorting to a number of built-in and customisable data entry forms, and (iv) *search and filter* information by using a powerful yet simple point-and-click mechanism that poses restrictions on the stored data and extracts the requested information in a number of formats and outputs including raw data, pie/column charts, and ready-to-process spreadsheets. Due to the cloud infrastructure, computational resources are allocated on demand, providing *elasticity* and *fault-tolerance*.

**Contributions.** The contributions of this work are twofold:

- We propose a *cloud-based, zero-cost, zero administration* tool that offers both fundamental and advanced user and data management functionality for multi-centre studies. To the best of our knowledge, this is the first cloud-based system that focuses on multi-centre studies and allows users to deploy their own data management platforms within minutes, alleviating the need to rely on expensive custom-made solutions that require IT infrastructure and skills to maintain.
- We present the architectural considerations and solutions behind the proposed tool, and propose a number of *novel services* that allow users without any prior IT knowledge to create, administer, launch, and use personalised data management platforms.

**Application Scenario.** As an example of an application scenario, let us consider three research groups (namely ABC, Bio, and Med) located in three geographically distributed hospitals and clinics. Group ABC is the coordinator group, i.e. the group that has decided to lead the multi-centre study and is responsible to set up the platform for the collection and processing of study data. Currently Mary (working for ABC), who is the person in charge of the setup of the data management platform, would need to get in touch with an IT company and explain the needs and specificities of the study. Subsequently, group ABC would need to buy and maintain a costly IT infrastructure onsite to host the developed solution,

A. Tsafara, C. Tryfonopoulos and S. Skiadopoulos are with the University of Peloponnese, Greece. {amtsafara,trifon,spiros}@uop.gr

and possibly hire an IT professional/company to keep both the system and the infrastructure up-to-date. Similar approaches need to be followed, even if Mary decides to resort on one of the free systems (e.g., REDCap [10]) offering data management services. Clearly, Mary and ABC would benefit from accessing a cloud-based service that would allow them to create and deploy such a platform in a fast, free, and effortless way. This system would be a valuable tool *beyond anything currently supported*, that would allow Mary and her group to save time, effort, and resources.

After the platform creation, Mary will be able to create and manage the users of the deployed platform, and define access control policies. These users will then be able to login and input data in the data management platform from any location or device. In our example scenario, Mary creates new users for the groups Bio and Med participating in the study. The person responsible for the Bio group goes through old patient records stored in the hospital archive and inserts them in the platform. At the same time, the person responsible for the Med group, visits a new patient and inserts the patient data in the system through his PDA. When the data collection phase is completed, the study coordinator may use the available searching and filtering techniques to issue appropriate queries and export data of interest for analysis (e.g., SPSS). Since the ABC group is coordinating the study, it has access to all inserted data, while other groups' access is restricted to the data policy enforced.

The rest of the paper is organised as follows. Section II discusses related work, while Section III introduces the system architecture and describes the implemented services and functionality. Finally, Section IV concludes the paper and discusses future research directions.

## II. Related Work

Over the years, many solutions aiming at the management and sharing of health/biomedical information have been proposed; we focus on approaches related to *electronic patient record (EPR)* systems, and discuss EPR systems specifically designed for multi-centre studies. EPR systems can speed up clinical communication, reduce the number of errors, and assist doctors in diagnosis and treatment. Many EPR systems (e.g., [4], [11], [16], [5]) aim at helping users to sort, archive, explore, export, and organise raw data like medical images scanned documents, laboratory data, and clinical ratings. Apart from data storage the offered functionality may also extend to quality control, and data analysis on the stored information to support clinical decision making [9] or promote medication adherence [14].

The aforementioned systems target research carried out at a *single site*, but EPRs have also been used in the context of *multi-centre studies*. Most of the proposed systems focus on a *specific study*, and put forward architectures and services tailored to the problem at hand. [15] presents an information system that may be used to manage multi-centre studies for cancer. Similarly, MSBase [7] introduces a web platform for collecting prospective data on patients with multiple sclerosis, while [6] presents a system for HIV/AIDS prevention and treatment. Finally, a number of EPR systems (e.g., [1], [3], [2], [8]) have also been designed focusing on the support of different types of multi-centre studies.

The large number of existing specialised systems and the needs dictated by each different multi-centre study led researchers to the design of systems that are able to support *classes* of functionalities needed in multi-centre studies. The most prominent paradigms in this line of work are the REDCap project [10] and the Qure system [12]. REDCap provides a principled way of designing, constructing, and managing databases that are able to support multi-centre studies. However, to deploy a system for a specific study, the study coordinator needs to get in touch with the REDCap project, and inform the IT specialist on the specific needs and requirements of the study at hand. Subsequently, after an iterative and possibly long process of refinement and corrections in the database design, the REDCap IT expert will deploy the database and the users will be able to enter the data. Obviously, any subsequent changes in the specifications will result in the redesign of the database and the porting of the inserted data in the new database. On the other hand, the Qure system, while supporting online creation of study questionnaires, offers (i) limited data types (e.g., does not offer custom drop-down lists or complex data types), (ii) no data export functionality (e.g., pie/column charts, spreadsheets, SPSS compatible output), (iii) a query engine with limited expressiveness, and (iv) runs on dedicated hardware with no adaptation policy (e.g., elasticity of resources) and low quality-of-service guarantees.

## III. System Overview

In this section, we outline the CloudStudyarchitecture, and present the associated services and functionality.

### A. Architecture

CloudStudy allows users to design and build data management platforms, through a series of simple and adaptive processes. This can be done transparently through simple-to-follow wizards from users without any IT training, while the cloud-based architecture automatically adapts to the required resources and infrastructure by relying on cloud elasticity.

The CloudStudy system has been entirely developed by *open source software*; it uses the Linux/Apache/PostgreSQL/PHP (LAMP) framework as the backend database infrastructure, while the rest of the modules have been developed using Javascript/PHP/JQuery. The cloud functionality is provided by the open-source platform ownCloud setup over a medium-sized computing infrastructure available at the University of Peloponnese. Figure 1 presents a high-level view of the system architecture and the different types of modules implemented. The Cloud API is responsible for performing all necessary communication with the ownCloud platform and provides elasticity services, while the DB Manager performs all necessary storage and retrieval operations to the database backend. The Study Manager module is responsible for the creation, editing, and management of studies, and consists of a number of modules utilised to (i) manage the users and the stored data associated with a study and (ii) filter/extract data requested by a user or a study administrator. The Platform Manager is used to create, edit, and manage platforms and templates utilised by different studies. The User Manager module is utilised by the system administrator to create and manage the study

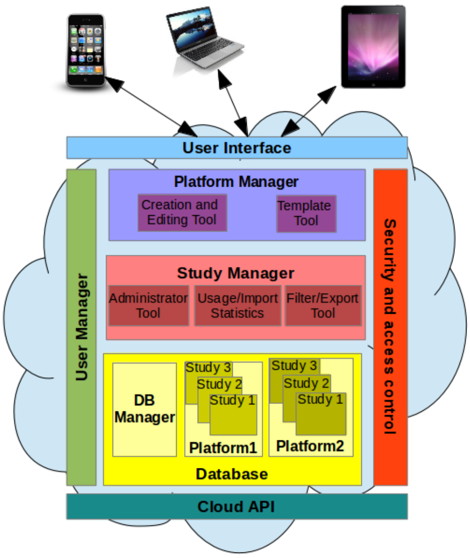
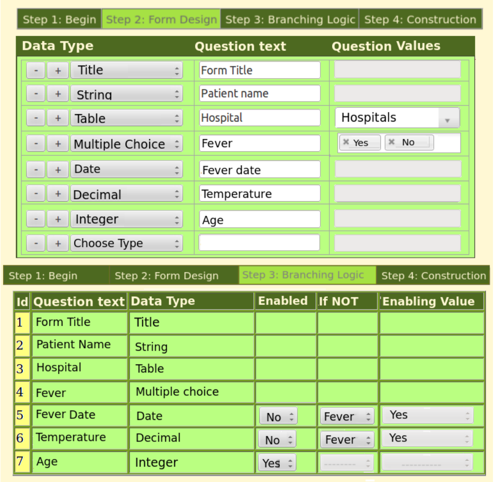


Fig. 1: A high-level view of CLOUDSTUDY architecture.

administrators (users in-charge of one or more studies), and also by the study administrators to create and manage the participants (and their roles) in a specific study. The Security and Access Control Module enforces the security policies for the system and controls access privileges over the stored data. Security features include certificate- and password-based authentication, single sign-on policy, and role-based user management. Finally, the User Interface module is responsible for identifying the hardware used to connect to CLOUDSTUDY (PC, tablet, smartphone) and adjust the viewing components accordingly.

### B. Creating platforms and platform templates

The platform creation and editing tool provides a flexible, adaptive and intuitive way to design a platform or a platform template. *Platforms* are custom-made questionnaires for each study together with all necessary user administration and data management components, while *platform templates* offer the ability to reuse all or part of a platform (e.g., the demographic data questionnaire) in more than one studies. The platform creation and editing tool guides the user to design a new platform by allowing him to specify a name for the platform/template and a number of study questions consisting of three columns: the *data type*, the *question text*, and the *question values* that depend on the data type of each question. The available data types are title (for introducing new questionnaire sections), string, integer, date, decimal, multiple choice, complex and table. More complex data types such as multimedia, time-series, and streaming data are not currently supported, as epidemiological studies usually focus on high-level descriptions of symptoms or diagnoses, and are not interested in details of medical examinations; however, in the future, we plan to offer the possibility of storing such data as reference points. The user may add, rearrange, or delete questions by dragging and dropping elements. Subsequently the user needs to determine the *branching logic* of the questions. This functionality allows the user to specify the values in other questions that are required to enable or disable following ones (e.g., if the answer to the question "Fever" is yes, enable the questions for the date and observed value). Figure 2 shows the construction of a questionnaire template with seven questions of different data types (top) and the branching logic menu (bottom). Editing an existing platform involves two different scenarios: (i) *textual editing* that refers to changes that involve rephrasing or addition of questions, or changing the branching logic of specific questions, and (ii) *structural editing* that refers to changes that may affect data consistency (e.g., question reorganisation/removal, modification of question datatypes) or may cause compatibility problems between the stored patient records and the new questionnaire. Both editing scenarios are supported in CLOUDSTUDY.



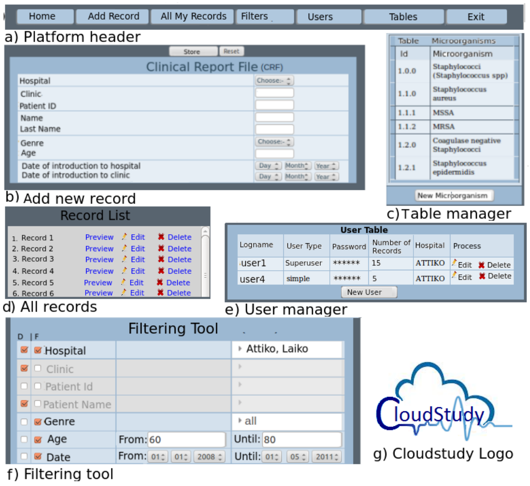
| Data Type | Question text | Question Values |
|---|---|---|
| - + Title | Form Title | |
| - + String | Patient name | |
| - + Table | Hospital | Hospitals |
| - + Multiple Choice | Fever | × Yes × No |
| - + Date | Fever date | |
| - + Decimal | Temperature | |
| - + Integer | Age | |
| - + Choose Type | | |

| Id | Question text | Data Type | Enabled | If NOT | Enabling Value |
|---|---|---|---|---|---|
| 1 | Form Title | Title | | | |
| 2 | Patient Name | String | | | |
| 3 | Hospital | Table | | | |
| 4 | Fever | Multiple choice | | | |
| 5 | Fever Date | Date | No | Fever | Yes |
| 6 | Temperature | Decimal | No | Fever | Yes |
| 7 | Age | Integer | Yes | -------- | --------- |

Fig. 2: Platform creation and editing.

To ensure data consistency across studies, and enhance data integrity and validation of input, CLOUDSTUDY provides users with the ability to dynamically create, store, and edit drop-down *lists of elements*. To do so the user specifies a unique name for the drop-down list and defines the list elements. Subsequently, when specifying a question, the user needs to set the data type to table –see Figure 2 (top)– and select one of the stored drop-down lists. Drop-down lists may also be deleted, given that no platform/template uses them.

To support complex medical operations such as therapeutic protocols, treatments, or antibiograms CLOUDSTUDY introduces the *complex data type* to model groups of recurring questions –as in treatment plans which usually consist of one or more antibiotics with periodically recorded data (e.g., name, start/end dates, outcome). The advantages of creating



a) Platform header

Clinical Report File (CRF)

| | |
|---|---|
| Hospital | Choose:- |
| Clinic | |
| Patient ID | |
| Name | |
| Last Name | |
| Genre | Choose:- |
| Age | |
| Date of introduction to hospital | Day Month Year |
| Date of introduction to clinic | Day Month Year |

b) Add new record

| Table | Microorganisms |
|---|---|
| Id | Microorganism |
| 1.0.0 | Staphylococci (Staphylococcus spp) |
| 1.1.0 | Staphylococcus aureus |
| 1.1.1 | MSSA |
| 1.1.2 | MRSA |
| 1.2.0 | Coagulase negative Staphylococci |
| 1.2.1 | Staphylococcus epidermidis |

c) Table manager

Record List

| | | | |
|---|---|---|---|
| 1. Record 1 | Preview | Edit | ✖ Delete |
| 2. Record 2 | Preview | Edit | ✖ Delete |
| 3. Record 3 | Preview | Edit | ✖ Delete |
| 4. Record 4 | Preview | Edit | ✖ Delete |
| 5. Record 5 | Preview | Edit | ✖ Delete |
| 6. Record 6 | Preview | Edit | ✖ Delete |

d) All records

User Table

| Logname | User Type | Password | Number of Records | Hospital | Process |
|---|---|---|---|---|---|
| user1 | Superuser | ****** | 15 | ATTIKO | Edit ✖ Delete |
| user4 | simple | ****** | 5 | ATTIKO | Edit ✖ Delete |

e) User manager

Filtering Tool

| D | F | | | |
|---|---|---|---|---|
| ☑ | ☑ | Hospital | | Attiko, Laiko |
| ☑ | □ | Clinic | | |
| □ | □ | Patient Id | | |
| ☑ | □ | Patient Name | | |
| □ | ☑ | Genre | | all |
| □ | ☑ | Age | From: 60 | Until: 80 |
| □ | ☑ | Date | From: 01 01 2008 | Until: 01 05 2011 |

f) Filtering tool

g) Cloudstudy Logo

Fig. 3: Administering an existing study.

complex data types include better modelling of the input data, and thus richer query possibilities and flexibility in the design of a study that involves complex medical processes.

### C. Administering and using a study

Once a study has been setup by creating a new platform, or re-using an existing platform template, the study administrator may use the study menu to perform a number of administrative tasks, that include *user management* and *role definition* (Figure 3(e)), *data input/correction/deletion* (Figures 3(b) and (d)), *data filtering/export* (Figure 3(f)), and generation of *usage/input statistics* (Figure 3(e)). Normal users may be created by the study administrator and their view of the data is limited to the access control policy defined. Data access privileges are defined in three layers: (i) normal users may access only the data input by them, (ii) study administrators may access all data in a study, and (iii) system administrators may access all data in the database. Users of a study may *add, edit, preview* and *delete* patient records (Figures 3(a), (b), and (c)) and are able to temporarily save a questionnaire and resume data input at another time.

To filter and query the stored data, CLOUDSTUDY (logo shown in Figure 3(g)) uses a powerful yet easy-to-use query issuing tool that allows users to filter and retrieve stored records by applying constraints with simple point-and-click interactions (Figure 3(f)). Using the tool involves a *two-step process*: in the first step the user is required to define the query output by checking the questions that will be used for *data projection* (i.e., data to be exported), while in the second step, the user applies one or more *filtering conditions* on the data. The filtering conditions are introduced by presenting the user *all distinct values* stored for a specific question and allowing him to define the ones that satisfy his filtering criteria. In this way, the user may define *conjunctions* and *disjunctions* both on the questions and on the stored data. The query result may then be stored in a spreadsheet, viewed as a pie/column chart, or as a list of records. The query issuing tool is able to capture complex data types, allowing the user to (i) set more than one filters for every complex data type and (ii) include constructs like concurrent episodes of a diagnosis or treatment.

## IV. CONCLUSIONS AND OUTLOOK

CLOUDSTUDY is currently under alpha testing for multi-centre studies led by the Hellenic Society for Chemotherapy and the University Hospital Attikon, and has already been used by more than 10 public hospitals in Greece. The preliminary user-feedback has been exploited to improve system functionality and design new services. Our plans include supporting more export formats and sophisticated data types, performing large-scale user studies to improve the usability of the user interface, and dealing with legacy issues by supporting heterogeneous data representations.

## REFERENCES

[1] CASCADE: Concerted Action on SeroConversion to AIDS and Death in Europe. Accessible at: http://www.ctu.mrc.ac.uk/cascade/.

[2] COBRED: Colon and Breast Cancer Diagnostics. Accessible at: http://www.cobred.eu/.

[3] HICDEP: HIV Cohorts Data Exchange Protocol. Accessible at: http://www.hicdep.org/.

[4] C.L. Adamson and A.G. Wood. DFBIdb: a software package for neuroimaging data management. *Neuroinformatics*, 2010.

[5] M. Arya, W. Cody, C. Faloutsos, J. Richardson, and A. Toga. QBISM: A prototype 3-D medical image database system. *IEEE Data Eng. Bull.*, 1993.

[6] A. Nucita et al. A global approach to the management of EMR (Electronic Medical Records) of patients with HIV/AIDS in Sub-Saharan Africa: the experience of DREAM Software. *BMC Medical Informatics and Decision Making*, 2009.

[7] H. Butzkueven et al. MSBase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Multiple Sclerosis*, 2006.

[8] H.S. Fraser et al. An information system and medical record to support HIV treatment in rural Haiti. *British Medical Journal*, 2004.

[9] S.J. Zasadaa et al. IMENSE: An e-infrastructure environment for patient specific multiscale data integration, modelling and clinical treatment. *Computational Science*, 2012.

[10] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde. Research electronic data capture (REDCap) A metadata-driven methodology and workflow process for providing translational research informatics. *Biomedical Informatics*, 2009.

[11] J.D. Van Horn and A.W. Toga. Is it time to re-prioritize neuroimaging databases and digital repositories? *Neuroimage*, 2009.

[12] M. Jager, L. Kamm, D. Krushevskaja, H. Talvik, J. Veldemann, A. Vilgota, and J. Vilo. Flexible database platform for biomedical research with multiple user interfaces and a universal query engine. In *DB&IS*, 2008.

[13] J. Lee. EMR management system for patient pulse data. *Med. Syst.*, 2012.

[14] M. Lopez-Nores, Y. Blanco-Fernandez, J.J. Pazos-Arias, and J. Garcia-Duque. The iCabiNET system: harnessing electronic health record standards from domestic and mobile devices to support better medication adherence. *Computer Standards and Interfaces archive*, 2012.

[15] M. Martinez, J.M. Vazquez, M.G. Lopez, F.M. Arnal, B. Gonzalez-Conde, J. Pereira, and A. Pazos. Semantic integration of data in an information system for multicenter epidemiological studies on cancer. In *MIE2008*, 2008.

[16] A. Pozamantir, H. Lee, J. Chapman, and I. Prohovnik. Web-based multi-center data management system for clinical neuroscience research. *Med. Syst.*, 2010.