Journal of Clinical Trials

Clivio et al., J Clin Trials 2016, 6:4
DOI: 10.4172/2167-0870.1000275

Research Article Open Access

Beyond the Web: A Decentralised Approach to Data Collection from Multi-Centric Clinical Trials by Virtual Organizations Using Peer-To-Peer Data Sharing

Luca Clivio[1], **Lital Hollander**[1], **Luca Beltrame**[1] **and Anthony J Travis**[2]

[1]*Mario Negri Institute for Pharmacological Research, Milan, Italy*

[2]*Minke Informatics Limited, Bridge of Alford, Scotland, UK*

[*]**Corresponding author:** Luca Clivio, Mario Negri Institute for Pharmacological Research, Milan, Italy, E-mail: luca.clivio@marionegri.it

Received date: June 30, 2016; **Accepted date:** July 28, 2016; **Published date:** August 2, 2016

Abstract

Harnessing the power of ‘disruptive’ technologies in a peer-to-peer network is a cost-effective way for non-profit research organizations to manage their clinical trials data and provide decentralised control to ensure clinical trials are conducted for the benefit of patients rather than for any one of the collaborating laboratories. Sharing data between multiple centres is important in clinical research to reduce any selection bias of patients participating in a trial when the patients involved at any one centre may not be representative of the population as a whole and the number of patients needed for a statistically meaningful result is greater than one centre can provide. Regulation of data sharing in clinical trials is essential to avoid accidental loss of data and to maintain control of access to data. Regular backups and measures to restrict unauthorised access to the data are an essential part of quality control for regulatory compliance of clinical data collection tasks. FDA CFR 21 part 11, which is an important regulation for clinical trials, explicitly require this level of compliance. Avoiding accidental 'prior-disclosure' of data and preventing unauthorised or inappropriate use of it is important for the correct attribution of work done by clinical research scientists. Many aspects of clinical research rely on ‘virtual organizations’ because management of clinical research projects often spans institutional boundaries to avoid duplication of effort and to share resources in order to solve problems economically. Clinical researchers working in different laboratories collaborate and share their data. The laboratories involved need to develop trust relationships to share data and intellectual property. This requires agreement about data management: In particular, where to store the shared data, which will curate it, who has access to it and how to share the data. In this paper we discuss a cost-effective and decentralised network-based approach to management of clinical data that disrupts established practice of using a centralised web-based database for all aspects of managing clinical trials data.

Introduction

A conventional business computing solution for data sharing in clinical research prioritises the protection of intellectual property and centralises resources in order to minimise the cost of the supporting ICT infrastructure. This approach also centralises control of access to the data with the partner providing the supporting ICT infrastructure. This requires an unlimited level of trust to be placed on the owner of the shared, centralised database by other collaborating partners who become clients. The justification often given to such “clients” about the importance of having a centralised database (usually a web based database) depends on their perceived advantage of several factors including: “Real time” access to information contributed by all partners in a collaborative network, no need for complex or expensive customised software on the client's computers that may require system administration rights to install, no upgrades of customised software need to be sent to the users, no need for stringent data safety or security policies on the client's computers and, in general, the fact that this represents well-established 'good practice'.

In contrast, a scientific computing solution for data sharing prioritises the sharing of intellectual property as a means to increase scientific knowledge and maximise the benefit of ICT to scientists. This requires a completely different strategy: A collaborative development of FLOSS (Free/Libre Open Source Software) for data sharing that provides researchers with the freedom to innovate and, if the software is reliable, for devolution of authority to access data to peer collaborators. Being reliable for software should not to be confused with its “it has been validated”. Devolution of authority for data access to collaborating peers allows, in principle, for multiple stakeholders to perform their own data validation.

The organization of ICT is often viewed from a business computing perspective and is usually considered to be a single cost centre. However, clinical research is a scientific activity and organising a data sharing network for clinical trials should the responsibility of the scientists involved, not that of a business IT manager. The objective of collaborating scientists who want to share data is to create a "self-organising" peer-to-peer network and, provided that privacy laws are complied with by all computer systems allowed connecting, sharing of scientific data in a peer-to-peer network is robust and avoids dependence on a third-party data centre that may represent a SPOF (Single Point of Failure).

A centralised web-based database for sharing clinical research data has several disadvantages from the end user's point of view including the requirement for: A permanent high-bandwidth network connection, limited or no possibility of working offline, a “minimally configured” web browser (i.e. no pop-up blockers, no java/ javascript/ cookies blockers, no add-ins) customised to the particular requirements of the web-based database. The illusion an end-user might have of a general purpose computer used to navigate the Internet and also used to access a web based database of clinical trials data is difficult to achieve in practice. Other, typical, drawbacks of web based databases for clinical trials data include being unable to have a local copy of the database without complex technical safety and security issues because these features are not implemented on the client-side of a web-based database and the difficulty of running "asynchronous tasks" where data is committed to the central servers, for example when executing an "omics" data query.

An important question is: Do we need yet another program to manage clinical trials data? If we do, is a decentralised peer-to-peer philosophy acceptable? Can we deal with many long or short temporary disconnections from a data sharing network? Is it acceptable to store a master copy of the data by a local collaborating centre without the costs and data privacy concerns of it being stored by a third party? Is it possible to store clinical data safely and securely locally without requiring administrative rights on the client's computer because these rights are most often denied to a clinical investigator when using a client computer owned by the research centre?

Materials and Methods

Installation of software requiring administrative rights on a client computer is usually considered to be the only alternative to using a web based data sharing application. However, the ICT policy at a typical hospital does not permit a clinical scientist to install software requiring administrative rights on a computer owned by the research centre. Despite this limitation, software viruses and malware/spyware work perfectly without administrative rights on computers maintained and managed by ICT staff. The truth is that no matter how strict the ICT policies of hospitals are, in many it is still possible to download and run an executable file with access to the network. An executable file for a MS Windows environment is typically an ".exe" file, but text documents and spread sheets or presentation files for the ubiquitous MS Office suite, and other software provided with scripting capabilities, are able to host executable code.

If the computers used by collaborating scientists who want to organise their clinical trial data are viewed as a resource owned by a virtual organization representing the collaboration, we can see that all the computers considered together are sufficiently powerful, well connected and have sufficient storage to offer adequate capabilities to perform all the clinical data collection, quality control, and storage redundancy for most of the clinical trials published at present. Our approach has been to build, in-house, open source software able to run without installation, using automatic discovery of its underlying network topology, auto-clustering of its peers discovered on the LAN (Local Area Network) hosting database deployments and able to share data without the use of a centralised database using, instead, a small number of index servers for the distributed database instances in any given geographic area.

Our strategy has been to employ similar techniques to those used in "spyware" to develop an autonomous agent and utilise computers in the virtual organization's network as a private cloud to implement a clinical data oriented "push-based peer-to-peer database" [1]. We have implemented the agent as a lightweight peer-to-peer servent (server + client) downloaded and run on a clinical researcher's own computer without installation or administrative rights. The agent is capable of updating itself automatically and able to work in a hybrid online/offline environment. The agent software is written in Python and is portable across three popular desktop computer platforms used by clinical researchers (Windows, Mac OSX, Linux). The initial download includes all the software required for handling clinical trials data, with packages for data analysis and on-demand connection to the peer-to-peer network. Asynchronous data transfer and software updates are implemented on-demand using a local repository for data storage and analysis. Shared network storage and computation facilities elsewhere are only used if needed and when the opportunity of connecting to the peer-to-peer network arises. There is no requirement for a central repository. The redundancy of multiple copies stored on the peer-nodes means that any peer available will be able to provide a source for the required information via the index servers if it does not have the information stored locally. The exact technology used is less important than the principle that even in a world of restrictive ICT policies of hospitals and research centres peer-to-peer data sharing is a possible using principle inspired by "malware" developers to provide clinical researchers with the means to share manage and interpret their clinical trials data in the way that they want to.

Our clinical data management system, HEAVyBASE, was developed from a previous web-based system GCPBASE [2], itself based on a hierarchical EAV (Entity Attribute Value) [3] data model by Chen provided with a custom push-based peer-to-peer transfer protocol based tunnelled into HTTP request in order to force full redundancy of all the data on all peers without being blocked by the common firewalls found in the centers. The technology used for developing this and details of how the decentralised data sharing is organised the beyond of the scope of this paper, but it represents one of many possible implementations of a decentralised platform for data sharing by virtual organizations.

Results

A decentralised platform for data sharing was built using the COTS (Commodity Off The Shelf) desktop computers already in use by the investigators involved in a clinical trial using open source software developed in-house and validated against FDA CFR 21 part 11 [4] for use in clinical trials, for collecting data from patients in a controlled randomised or observational clinical trial and in a retrospective registry and a Biobank. HEAVyBASE has been categorised as GAMP Category 5 (bespoke software) for atypical use of computational and network capabilities. The validation process was completed in 2014 with an audit of the ECRIN (European Clinical Research Infrastructure Network) in the Laboratories of Clinical Research and Life Science Informatics of the Oncology Department at the Mario Negri Institute for Pharmacological Research [5], where this data management philosophy and software were created. There are currently 25 controlled clinical trials and bio banks using this technology at the Mario Negri Institute, as promoter, and another 50 clinical trials and registries, with about 30,000 patients enrolled in various research institutes around Europe, and it has been clearly demonstrated that most of the hospitals in Europe have, at least, the potential for utilising the otherwise unused computational resources of the investigators who want to be involved in a clinical trial with few, if any, configuration issues or special permission given to them by the ICT manager of the centre, simply using the "freedom" already given to normal computer users to do their work with hospital-provided computers.

Discussion

Using a completely decentralised platform allows access to data (and funding) to be given only to the investigators actually involved in a clinical trial, and no one else, because no additional ICT infrastructure is needed. This is quite different to the data access and capital funding required for a centralised platform. The decentralised approach does, however, have certain drawbacks: In particular, information is first passed between "live peers" (i.e peers that are connected to the network) and it is not possible to have synchronous data alignment among all peers, which means that the use of a decentralised method for sharing data must be used with more awareness of its limitations than when using a completely synchronous system. A similar limitation can occur with randomization with an RCT (Randomised Controlled Trial). This type of system can be used under certain conditions: Even if a centralised randomization service is available, it is important to understand that stratification by Institute of the random list or dynamic randomization algorithm often happens because randomisation will be influenced by missing data that has not yet arrived from a participating centre or if previous randomizations have been done at another centre and the information about that randomisation has not yet arrived.

The approach adopted at the Mario Negri Institute is to have a completely decentralised data sharing engine, but with a centralised mail server to sending random notifications to all the stakeholders on every randomisation. This allows a centralised random counter to be obtained over stratification factors with a small overhead of having to maintain a mail-server but no clinical data is collected on the mail server, otherwise stratification factors would be incorrect, ensuring this way a correct number of assigned centres balance. Another limitation of the decentralised platform is data monitoring activity, which is basically a centralised activity. This type of decentralised approach involves having a full copy of all the data on every peer node (even if every user can access only his own sub-set of the data), but since the clinical monitor can be registered as one of the peer nodes, he can access all the data to be monitored at any time provided that the data to be checked have already been pushed to his own copy. For this reason, in our implementation, we provide a way to immediately import data from a centre to be monitored by allowing the dataset to be re-encrypted and sent as a single package via a dedicated file sharing system.

Periodic data review is another important activity to be considered: The appropriate way to perform any kind of edit-check must be implemented on the client-side because of the nature of this type of data collection. Every peer must be able to control itself, because a remote centralised data quality check does not necessarily work on a fully updated dataset on an asynchronous platform. The same file sharing technique described above can be used for ad interim or final analysis to be sure that it is performed on the final dataset. Last but not least, is that a clinical investigator needs to understand that, in a collaborative distributed peer-to-peer database, if a partner updates data without allowing time for the network to share the modifications and turns the system off, those modifications will only be on his computer until he turns it on again for a long enough time for the peer-to-peer agent to connect to the network and share his data. In general the system works like a telephone network or a chat system: you can receive messages only while your terminal is turned on. If this limitation is taken in account, this type of approach is probably the most cost-effective way to obtain a good quality dataset from a clinical trial.

References

1. Perez de Laborda C, Popfinger C (2004) A Flexible Architecture for a Push-based P2P Database D´ıgame Architecture. In 16th GI-Workshop on the Foundations of Databases, Universitat Dusseldorf.
2. Clivio L, Tinazzi A, Mangano S, Santoro E (2006) The contribution of information technology: Towards a better clinical data management. Drug Dev Res 67: 245-250.
3. Chen RS, Nadkarni P, Marenco L, Levin F, Erdos J, et al. (2000) Exploring performance issues for a clinical database organized using an entity-attribute-value representation. J Am Med Inform Assoc 5: 475-487.
4. Rockville MD (2003) 21 CFR Part 11. Electronic Records; Electronic Signatures; Final Rule. Food and Drug Administration.
5. Canham S, Clivio L, Cornu C, Crocombe W, De Bremaeker Nancy, et al. (2012) Requirements for Certification of ECRIN Data Centres, with Explanation and Elaboration of Standards.