

**SHORT COMMUNICATION****Efficient, quality-assured data capture in operational research through innovative use of open-access technology**A. M. V. Kumar,¹ B. Naik,² D. K. Guddemane,² P. Bhat,² N. Wilson,¹ A. N. Sreenivas,² J. M. Lauritsen,^{3,4} H. L. Rieder^{5,6}<http://dx.doi.org/10.5588/pha.13.0004>

Ensuring quality of data during electronic data capture has been one of the most neglected components of operational research. Multicentre studies are also challenged with issues about logistics of travel, training, supervision, monitoring and troubleshooting support. Allocating resources to these issues can pose a significant bottleneck for operational research in resource-limited settings. In this article, we describe an innovative and efficient way of coordinating data capture in multicentre operational research using a combination of three open access technologies—EpiData for data capture, Dropbox for sharing files and TeamViewer for providing remote support.

Quality-assured data entry has been aptly described as the ‘Cinderella’ of medical research.¹ Although there are several ways of reducing data entry errors, double data entry and validation has been considered the definitive gold standard,² whereby data are entered independently twice and value pairs compared for discordances, followed by resolution of discordances by referral to the original data source.³ However, a review of all publications in the *International Journal of Tuberculosis and Lung Disease* in the year 2008 indicated that only 2/43 published articles related to tuberculosis (TB) epidemiology actually mentioned achieving this standard, while more than half did not even mention data quality.⁴ This tendency to unquestioningly presume that data in published research is of high quality is highly questionable. Despite the easy availability of high-quality, open-access tools for quality-assured data capture, the concept has apparently been grossly neglected in academic curricula in medical and public health schools.

The challenge of assuring data quality is compounded if the context is a multicentre research study involving multiple study sites and personnel, where the costs of travel for training, supervision, monitoring and troubleshooting support are often substantial. These costs can become a significant barrier to assuring data quality while conducting high-quality operational research, especially in resource-limited settings. In this article, we describe an innovative model that used multiple open-access tools in a multicentre operational research project with efficient use of resources.

CONTEXT

The subject of this operational research project was the implementation of provider-initiated human im-

munodeficiency virus (HIV) counselling and testing for patients with presumptive TB under routine programmatic conditions in the State of Karnataka, India. This State has a population of about 62 million, and is divided into 31 districts spread across ~192 000 km², extending 750 km from north to south and 400 km from east to west. The study took place between January and March 2012.

As part of the intervention, every patient with presumptive TB attending the microscopy centre was assessed for HIV status by the laboratory technician. Those with unknown HIV status were referred to the nearest HIV testing facility. HIV test results were captured by trained staff in a structured paper-based data collection form with measures built in to ensure data validity. Once the data were on paper, the forms were brought to the district TB centre for compilation and electronic capture. This meant that data entry would occur at 31 sites across the state, and the responsibility was assigned to the data entry operator of the district under the Revised National Tuberculosis Control Programme (RNTCP).

INNOVATION: SETTING UP THE DATA CAPTURE SYSTEM FOR EACH SITE

We used three tools to set up the system for data entry.

- 1 First, we used EpiData Entry version 3.1 (EpiData Association, Odense, Denmark, <http://www.epidata.dk>) to design the data capture instrument. In addition to being open access, this tool offers several advantages: small size of the software and data files, non-interference of the software with the operating system of the user's computer during installation, excellent capabilities for inbuilt checks during data entry to reduce the frequency of data entry errors, and a simple option that allows double data entry and validation. Most importantly, it is user friendly, making it easy to teach and learn.
- 2 Second, we used Dropbox (<http://www.dropbox.com>) to share the folders over the web (Box). The principal investigator (PI) designed the data capture formats in EpiData and placed them in a district-specific Dropbox folder (31 folders, one for each district), and an invitation was sent by e-mail to each district data entry operator to join the shared folder. Once the user had accepted the invitation and installed Dropbox, the software created the

AFFILIATIONS

- 1 International Union Against Tuberculosis and Lung Disease (The Union), South-East Asia Regional Office, New Delhi, India
- 2 World Health Organization Country Office for India, New Delhi, India
- 3 Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark
- 4 EpiData Association, Odense, Denmark
- 5 The Union, Paris, France
- 6 Institute of Social and Preventive Medicine, University of Zurich, Switzerland

CORRESPONDENCE

Dr Ajay M V Kumar
International Union Against Tuberculosis and Lung Disease
The Union South-East Asia Office
C-6, Qutub Institutional Area
New Delhi, India 110016
e-mail: akumar@theunion.org; sathyasaakshi@gmail.com

ACKNOWLEDGEMENTS

The authors thank the data entry operators working for the Revised National Tuberculosis Control Programme in the State of Karnataka, who played a key role in implementing this innovative model. The publication of this research was supported through an operational research course that was jointly developed and run by the Centre for Operational Research, International Union Against Tuberculosis and Lung Disease (The Union); The Union South-East Asia Office, New Delhi, India; and the Operational Research Unit, Médecins Sans Frontières, Brussels Operational Centre, Luxembourg. Funding for this course came from an anonymous donor and the Department for International Development, UK. Conflict of interest: none declared.

KEY WORDS

EpiData; quality-assured data capture; data validation; India

Received 11 January 2013

Accepted 14 January 2013

What is Dropbox?

Dropbox is a file hosting service operated by Dropbox Inc. that offers cloud storage and file synchronization. Dropbox uses a 'Freemium' business model, where users are offered a free account with a set storage size (2 gigabytes in this case) and paid subscriptions for accounts with more capacity. Dropbox allows users to create a special folder on each of their computers, which Dropbox then synchronises so that it appears to be the same folder (with the same contents) regardless of the computer it is viewed on. Files placed in this folder are also accessible through a website and mobile phone applications. Such folders can be shared with others for mutual access. More information at www.dropbox.com

What is TeamViewer?

TeamViewer is a secure software package for remote control, desktop sharing, online meetings, web conferencing and file transfer between computers. TeamViewer is a tool that makes it very easy to set up and use a Virtual Private Network connection that lets you take complete control of another computer from your own computer via internet. It enables two-way connections in which users can flip control back and forth. While TeamViewer is proprietary, it is free for non-commercial purposes. More information at www.teamviewer.com

shared folder on their local computer. This provided the powerful, simultaneous option of both offline data entry and online file synchronisation. The option of offline data entry ensured that continuous internet connectivity was not required during data entry. Online synchronisation meant near real-time sharing of data with the PI and data safety through online backup. Any need to change the structure of the data capture instrument to suit the needs of an individual site could be achieved very easily by manipulating the files in the shared folder by the PI. Without this instrument, a physical visit would have been required or considerable time expended in e-mailing files back and forth.

- 3 Third, we used TeamViewer (<http://teamviewer.com>) to provide support remotely to the individual sites for initial setup and troubleshooting (Box). We used this software to connect to the district computers and set up the data capture system, including software installation. This was also utilised as a training opportunity. After establishing telephone contact, the PI explained and demonstrated the use of the data capture system. The user was then allowed to enter the data from a few records under PI supervision, allowing resolution of any early problems. Thereafter, if the user encountered any problem during data entry, he or she would connect with the PI on TeamViewer to show the error to the PI and have it resolved almost immediately. A problem that hitherto required a visit in person to resolve could now be solved very easily and efficiently during a TeamViewer session.

Data entry

Once the system for data capture was set up, the data entry operators entered the data twice into files designated for the purpose and already placed in the shared Dropbox folder. When they had finished the data entry, this was communicated to the PI by e-mail. The PI then performed a 'data validation' ('data compare') between

the two databases and generated a validation report listing discrepancies between the two databases, placing the report in the Dropbox folder. The data entry operator was informed and requested to refer to the original data for correction and finalisation.

What was achieved? What resources were required?

Over a span of 3 months, data on nearly 115 000 study participants were electronically captured while ensuring the highest standards of data quality coordinated across 31 sites. The PI was assisted by three colleagues working for the RNTCP in the state in ensuring that the data capture system was set up in their respective districts. These three colleagues also had access to the shared folders on Dropbox. They together co-ordinated data capture in a period of 3 months over and above their routine job responsibilities.

DISCUSSION

This is an innovative use of open-access technology to coordinate data capture in a multicentre operational research project across 31 sites. While internet connection is decisive for success, the system does not require a large bandwidth nor 'always-on' capability: as EpiData file sizes are small (1000 records required only 216 kilobytes), bandwidth is of minor importance; as Dropbox has the option of offline access to shared folders, uninterrupted internet connection is not required for data entry, a major prerequisite in web-based data capture systems. Telephone costs can be circumvented, if required, using Skype™ (<http://www.skype.com>), which was used for periodic video conference calls between the PI and co-investigators to monitor the project. A recent study has shown that an alternative to double entry could be 'Automatic Forms Processing', a method by which one can 'automatically' capture information entered into data fields by scanning, and converting it into an electronic format through techniques such as 'Optical Mark Recognition' or 'Intelligent Character Recognition',⁵ but this is applicable only in highly structured questionnaires with only check boxes and numbers and no dates.⁵ This would also require relatively expensive equipment and computer expertise that are often not available in resource-limited settings.

Overall, this model for data capture proved highly efficient in terms of optimum utilisation of resources, including time, and we feel that it can be easily replicated in any resource-limited setting for operational research.

References

- 1 Cartwright A, Seale C. The natural history of a survey: account of the methodological issues encountered in a study of life before death. London, UK: King Edward's Hospital Fund, 1990.
- 2 Ohmann C, Kuchinke W, Canham S, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. *Trials* 2011; 12: 85.
- 3 Rieder H L, Lauritsen J M. Quality assurance of data: ensuring that numbers reflect operational definitions and contain real measurements. [State of the Art Series. Operational Research. Number 3 in the series]. *Int J Tuberc Lung Dis* 2011; 15: 296–304.
- 4 Rieder H L. What knowledge did we gain through *The International Journal of Tuberculosis and Lung Disease* in 2008 on the epidemiology of tuberculosis? *Int J Tuberc Lung Dis* 2009; 13: 1219–1223.
- 5 Paulsen A, Overgaard S, Lauritsen J M. Quality of data entry using single entry, double entry and automated forms processing--an example based on a study of patient-reported outcomes. *PLoS ONE* 2012; 7: e35087.

La garantie de la qualité des données au cours de la capture des données électroniques a constitué une des composantes les plus négligées de la recherche opérationnelle. De plus, les défis des études multicentriques comportent des problèmes concernant les logistiques du voyage, la formation, la supervision, le suivi et le soutien au dépannage. Une allocation de ressources à ces problèmes peut constituer un goulot significatif pour les recherches opérationnelles dans les

Garantizar la calidad de los datos durante la captura electrónica ha sido uno de los componentes más desatendidos de la investigación operativa. Los estudios multicéntricos afrontan además dificultades relacionadas con la logística del desplazamiento, la capacitación, la supervisión el seguimiento y de la asistencia técnica para la resolución de los problemas. La atribución de recursos a estas esferas puede representar un cuello de botella de la investigación operativa en los en-

contextes à ressources limitées. Dans cet article, nous décrivons une façon innovatrice et efficace pour la capture des données de coordination dans la recherche opérationnelle multicentrique au moyen d'une combinaison de trois technologies libres d'accès—EpiData pour la capture des données, Dropbox pour le partage des dossiers et TeamViewer pour procurer un soutien à distance.

tornos con escasos recursos. En el presente artículo se describe un método innovador y eficaz de coordinar la captura de datos en investigación operativa mediante la aplicación de tres tecnologías de acceso abierto, a saber: EpiData en la captura de datos, Dropbox en la puesta en común de los archivos y TeamViewer en la prestación de asistencia técnica a distancia.