Short EpiData course: do participants use the data entry tool two years post-training?

A. M. V. Kumar,¹ P. Chinnakali,² H. Shewade,¹ V. Gupta,¹ P. Nagpal,¹ A. D. Harries^{3,4}

<http://dx.doi.org/10.5588/pha.15.0034>

Setting: Training courses on data entry are few and far between compared to analysis. To address this gap, we conducted a short 2-day course on quality-assured data capture using EpiData for public health professionals in Bangalore and Puducherry, India, in 2013.

Objective: To determine the proportion of participants who used EpiData and taught the software to others during the 2-years after training and explore the reasons for doing/not doing so.

Design: Cross-sectional study with self-administered, semi-structured questionnaire developed using web-based Google Forms. We performed a manual thematic analysis to identify the major reasons for using/not using EpiData.

Results: Of 46 participants, 38 (83%) responded. Of 31 participants involved in research, 17 (55%) had used EpiData, of whom 6 (35%) had performed double entry and validation. Of the 14 who did not use EpiData, 11 had used MS Excel or SPSS/Epi Info for data entry. Of the 38 respondents, 29 (76%) had taught EpiData to other colleagues and students. Reasons for using EpiData included its user-friendliness, its being open access and the ease in preventing data entry errors. Reasons for not performing double entry included lack of time and manpower.

Conclusion: The short course on EpiData was effective in knowledge transfer and provides a scalable model for incorporation into the teaching curricula of medical schools and research institutions.

Research work involves data collection which, if performed using paper-based forms or registers, needs to be entered in an electronic database for easy analysis. Errors commonly occur during data entry, and ensuring quality during data entry is one of the most neglected components of research.^{1,2} There are several ways of reducing errors in data entry, including working together, whereby one person reads out the data and another enters the data by repeating it aloud, or using inbuilt checks within the data entry software that prevent the entry of implausible values. Double data entry and validation is considered the gold standard in minimising data entry errors; with this method, data are entered independently twice, and value pairs are then compared for discordances, followed by resolution of discordances by referral to the original data source.^{3–6}

While there are many tools available for analysis, very few ensure quality-assured data entry. Some of the high-quality data entry tools available include Epi-

Data (EpiData Association, Odense, Denmark), Epi Info™ (US Centers for Disease Control and Prevention, Atlanta, GA, USA) and CPro (US Census Bureau, Suitland, MD, USA), all of which are open access. Although the availability of open-access tools is a pre-requisite for progress, it is not sufficient. Despite the availability of these tools, quality-assured data capture has apparently been ignored in many published journal articles and academic curricula in medical and public health schools.⁷ Anecdotal evidence suggests that while many training programmes have focused on data analysis, few have focused on data entry.

To address this gap in capacity, the International Union Against Tuberculosis and Lung Disease (The Union) South-East Asia Office, in collaboration with the Bangalore Medical College and Research Institute, Bangalore, and the Jawaharlal Nehru Institute of Medical Education and Research, Puducherry, conducted two short courses in 2013 on quality-assured data capture using EpiData. The courses were targeted towards public health professionals from medical colleges and other research institutions working in close collaboration with India's Revised National Tuberculosis Control Programme.

While most of the participants reported in their end of course feedback that it was a useful learning experience, the ultimate success of a capacity building initiative is judged by whether participants actually use the tool in their research work and help to build the capacity of others. In this study, we aimed to determine the effect of the EpiData course on the participants' work 2 years after the training course. The specific objectives were to determine 1) the number (proportion) of participants who had used EpiData for data entry in at least one research project; 2) the number (proportion) of participants who had trained others to use EpiData and the number trained; and 3) reasons for using/not using EpiData.

METHODS

Study design

This was a descriptive, cross-sectional study with a one-time assessment 2 years after the course using a self-administered, semi-structured questionnaire.

Setting

The short 2-day course was developed and conducted by The Union South-East Asia Office, with funding support from the Eli Lilly Foundation (Indianapolis, IN, USA). From the perspective of the participants, the

AFFILIATIONS

- 1 International Union Against Tuberculosis and Lung Disease (The Union), South-East Asia Office, New Delhi, India
- 2 Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India
- 3 The Union, Paris, France
- 4 London School of Hygiene & Tropical Medicine, London, UK

CORRESPONDENCE

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

ACKNOWLEDGEMENTS

We would like to thank N Wilson, former regional director of The Union, South-East Asia Office, for his visionary leadership in developing an OR unit in India and being instrumental in initiating these short courses. We would also like to thank H L Rieder (<http://www.tbrieder.org/>), who has been a great advocate of quality-assured data capture and analysis using open-access tools in operational research and has kindly guided me (AMVK) through this journey. Special thanks are due to J M Lauritsen, founder of EpiData Association in Denmark, for his guidance and mentorship on several occasions. Finally, we thank all our partners, donors and other stakeholders who have chosen to collaborate with us in planning and conducting the short courses. Conflicts of interest: none declared.

KEY WORDS

quality-assured data entry; EpiData; data validation; India; capacity building

Received 24 June 2015
Accepted 18 September 2015
First published online
12 November 2015

PHA2015;5(4):261–265
© 2015 The Union

objectives of the course were 1) to understand the importance of quality-assured data capture; 2) to have a working knowledge of basic EpiData software; and 3) to use this knowledge to develop relevant data entry tools.

In this 2-day course, teaching involved lectures and software demonstrations followed by exercises developed using a real operational research project. All participants who attended the course were nominated by the heads of the respective institutions. Participants were given an overview of EpiData entry (version 3.1), and were taught how to make a data entry instrument with in-built checks and perform double data entry and validation through exercises. Trained mentors helped the participants during hands-on sessions during the 2-day workshop.

At the end of the workshop, participants were given an individual assignment to apply the lessons learnt by developing a data entry instrument for a given research protocol within a 2-week period. Mentors guided the participants in the assignments where necessary. On successful completion of their assignment, participants received a certificate of completion.

The course content may be accessed from <https://goo.gl/JPVke3>. The course was video-captured, edited and posted on YouTube (<http://goo.gl/3DdVSv>) for wider reach.

Study population and study period

All participants who successfully completed the two short courses conducted in Bangalore and Puducherry, India, in 2013 were included in the study.

Data collection

Data were collected between April and June 2015 using a pre-tested, semi-structured questionnaire with closed and open-ended questions (Table). The questionnaire was developed using an

open-access, web-based survey tool, Google Forms (Google Inc., Mountain View, CA, USA), and the web link to access the questionnaire was e-mailed to the participants. Those who did not respond were sent reminders by e-mail and/or telephone on at least three occasions.

Data analysis

Quantitative data were analysed using EpiData (version 2.2.2.183), and summarised using frequencies and proportions. Thematic analysis was used to analyse the open-ended responses for questions related to using/not using EpiData and/or double entry and validation. The data obtained were compiled, and the principal investigator read the responses to become familiar with the data. To ensure that the results were a reflection of the data, the themes were related back to the original data. The analysis was reviewed by a second reviewer to reduce bias and enhance interpretive credibility. Given the small amount of data, we performed the analysis manually.

Ethics

The Ethics Advisory Group of the International Union Against Tuberculosis and Lung Disease, Paris, France, determined that ethics clearance was not required for this study.

RESULTS

Of 46 participants from 18 institutions who successfully completed the course, 38 (83%) (29 [76%] males; median [range] age 33 [25–56] years) responded to the questionnaire. Of these, 24 (65%) were medical college faculty and the rest included postgraduate medical students ($n = 7$), scientists at research institutes ($n = 2$), a medical records technician ($n = 1$), a research

TABLE Questionnaire used for data capture to assess the impact of short EpiData courses, India, 2015

- 1 Date of birth (dd/mm/yyyy)
- 2 Sex (Male/Female/Transgender/Other)
- 3 Job designation when on the course
- 4 Current job designation and institute
- 5 Contact phone number
- 6 Did you use EpiData before the course? (Yes/No)
- 7 Have you been part of any research project (include student projects as well as MD/MS dissertations) since the course (Yes/No).
If no, go to question 11, otherwise go to the next question.
- 8 Have you used EpiData for at least one research project after the course? (Yes/No)
- 9 If you answered 'yes' to question 8:
 - a Provide details of the projects (include student projects as well as MD/MS dissertations)
 - b What made you choose EpiData over other software?
 - c Did you do double entry and validation in a research project? (Yes/No/NA)
 - d If no, describe the reasons
- 10 If you answered 'no' to question 8:
 - a Which software was used for data entry?
 - b Reasons for not using EpiData
- 11 Did you use EpiData for any other purpose? (capture of routine health information/ surveillance/registers in health centres, etc) (Yes/No)
 - a If yes, provide details
- 12 Did you teach/train your colleagues/students in EpiData? (either formally or informally) (Yes/No)
- 13 If you answered 'yes' to question 12:
 - a Did you make use of course materials (manual, PowerPoint) provided during training? (Yes/No/NA)
 - b Provide details of who was trained and how many (approximate numbers)
 - c Of the above, how many are using EpiData in their work?
- 14 If you answered 'no' to question 12, mention reasons for the same
- 15 Are you interested in facilitating future EpiData courses conducted by The Union? (Yes/No)
- 16 Please use this space for any other remarks you want to make

MD = Doctor of Medicine; MS = Master of Science; NA = not applicable.

associate in a non-governmental organisation ($n = 1$) and a data analyst ($n = 1$).

Of the 38 participants, 35 (92%) did not know EpiData before the course and were learning about it for the first time. During the 2 years after training, 31 (82%) participants participated in at least one research project (median 2 [range 1–6]), of whom 17 (55%) used EpiData.

The key reasons for choosing EpiData can be categorised into the following themes: simplicity and ease of use; free availability; the ability to reduce the possibility of error and ensure data quality with inbuilt checks and the unique feature of double entry and validation; the ability to enable quicker and uniform data entry across multiple sites, especially in multicentre research studies; smaller file size, enabling easy data sharing by e-mail among colleagues; the stable and interactive nature of the software; minimal need for or reduced time for data cleaning before analysis; and ease of training others in data entry.

Of those who used EpiData, 6 (35%) performed double entry and validation. The main reasons for not performing double entry included lack of time, insufficient manpower and large sample size, although two participants mentioned that they were determined to perform double entry in their next project.

Of 14 participants who did not use EpiData, 11 used Microsoft Excel (Microsoft Corp., Redmond, WA, USA) and the rest used SPSS (IBM Corp., Armonk, New York, USA) and/or Epi Info for data entry. One person mentioned that they had used Epi Info before the course but changed to EpiData after the course. The main reason for not using EpiData was increased familiarity and comfort level in using other tools, not only for themselves but also for their research team members and other departmental colleagues. One person mentioned conducting a research project a year after the course, by which time he had forgotten how to use EpiData and did not have time to refresh quickly. Another mentioned that the sample size was too small to warrant the effort required to set up an EpiData database, which was perceived to be tedious and time consuming.

Eight participants said that they had used EpiData for purposes other than research, which included conducting injury surveillance in the emergency department of a hospital, population enumeration in areas served by their urban health centre, hospital census, setting up community health information systems, analysis of course feedback and as part of undergraduate and postgraduate teaching curricula on data management.

A total of 29 (76%) participants said that they had taught EpiData to other colleagues and students in their settings, and 25 (86%) had used the course materials shared by The Union in the course. The median (IQR) number of people trained was 6 (3–12), and ranged from 1 to 130. Five participants mentioned having trained more than 50 participants, each of whom included undergraduate and postgraduate medical students, faculty and people working with national programmes. Six participants had already co-facilitated courses conducted by The Union and two thirds mentioned that they were willing to co-facilitate in future Union courses.

DISCUSSION

This survey showed that about half of the participants had used EpiData in their research projects and three quarters had taught the tool to others. This is encouraging, given the poor background of medical researchers in using software in general and the fact that most of the participants were not aware of EpiData

before the course. By thinking strategically about the selection of participants, and with adequate mechanisms for providing follow-up support, this output can be improved further. Although the users appreciated the software for its simplicity and user friendliness, only a third performed double entry and validation, which is considered the gold standard in assuring quality, due to lack of time and dedicated manpower for data entry. This is an important lesson, underlining the importance of dedicated budgets for quality-assured data entry in research projects. EpiData software also helps in calculating data entry time and estimating the human resource requirements accordingly.⁸ Some participants mentioned the usefulness of EpiData in multicentre research studies, as has been shown previously.⁹

Most of the participants who did not use EpiData reported using MS Excel or SPSS for data entry. These tools are primarily designed for data analysis, and are limited in their ability to ensure quality during data entry. One reason for the continued use of other tools was the comfort level and familiarity associated with them and the general resistance towards learning a new tool. This indirectly indicates the continuing neglect of the importance of data quality in research in general and in medical schools in particular. Alternatives other than familiarity or practice may act as a barrier to the comprehensive assurance of data quality. This can be overcome to a certain extent, and several measures are required to address this neglect. First, data entry should be formally integrated into the teaching curriculum, and all students conducting research should be mandated to use a tool to assure data quality. Second, assurance of data quality during data entry should be explicitly mentioned in the reporting guidelines for published literature, such as the CONSolidated Standards of Reporting Trials (CONSORT), the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidelines.^{10–12} Third, journals should require authors to include a statement on data quality in their submitted manuscripts.

Some participants mentioned that they had forgotten the steps in creating the data entry files, which was perceived to be tedious and time consuming, thus compromising data quality. This is expected to happen if there is not frequent revision in the initial learning phase, and we plan to mitigate this problem by providing teaching aids (such as videos) for easy revision in future courses. A new version of EpiData is now available (<http://epi-data.dk/download.php>) that has enhanced user-friendliness and reduces the initial learning curve.

It was encouraging to note that EpiData was used for purposes other than research, including population enumeration and setting up disease surveillance and health information systems in the primary health centres. Changes such as these are likely to strengthen health systems in general and increase overall efficiency. Data obtained through such systems are likely to be quality-assured and this bodes well for conducting future operational research.¹³

About three quarters of the participants taught the tool to several colleagues and students in their network. This indicates the easy scalability of short courses, strengthened by standardised course content and teaching aids. Two participants mentioned having taught EpiData to hundreds of students in their institutions, possibly indicating that they had successfully integrated EpiData into the teaching curriculum. To enhance this process, we plan to groom the interested participants as facilitators by providing them with opportunities for teaching and mentorship in the courses conducted by The Union over the next 5 years.

As most of the participants had not heard about the EpiData tool before the course, there is a clear need for advocacy to increase the awareness of EpiData among medical researchers. Pre-conference workshops at national and regional level and meetings of professional associations may be utilised for this purpose. Self-learning videos posted online are also helpful in increasing awareness.

The study has some strengths and some limitations. First, to the best of our knowledge, this is the first study to assess the impact of a training course on data entry. Second, the study assessed long-term impact, including not only knowledge transfer but also use of knowledge. This indicates sustainable, long-term impact of the course beyond that indicated by the immediate post-course feedback. Third, the use of Google Forms for data collection reduced desirable responses to some extent due to the absence of face-to-face interactions; however, some social desirability bias cannot be ruled out given that the information was self-reported and could not be validated. Future assessments could incorporate some methods of validating the self-reported information by asking participants to share a copy of the EpiData triplet files used for data capture. Limitations are related to the non-response by 15% of the participants and the small sample size, which precluded us from performing a subgroup analysis. Finally, it was not possible to tease out whether the observed effect was due to the training course or due to the EpiData tool itself.

In conclusion, the short, 2-day course on EpiData entry was effective in knowledge transfer and knowledge use and provides a scalable model for incorporation into the teaching curricula of medical schools and research institutions.

References

- 1 Rieder H L, Lauritsen J M. Quality assurance of data: ensuring that numbers reflect operational definitions and contain real measurements. *Int J Tuberc Lung Dis* 2011; 15: 296–304.
- 2 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.
- 3 Ohmann C, Kuchinke W, Canham S, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. *Trials* 2011; 12: 85.
- 4 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.
- 5 Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. *Control Clin Trials* 2003; 24: 560–569.
- 6 Atkinson I. Accuracy of data transfer: double data entry and estimating levels of error. *J Clin Nurs* 2012; 21: 2730–2735.
- 7 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.
- 8 Hoa N B, Chay S, Chen W, Lauritsen J M, Rieder H L. Human resource requirements for quality-assured electronic data capture of the tuberculosis case register. *BMC Res Notes* 2012; 5: 75.
- 9 Kumar A M V, Naik B, Guddemane D K, et al. Efficient, quality-assured data capture in operational research through innovative use of open-access technology. *Public Health Action* 2013; 3: 60–62.
- 10 Vandembroucke J P, von Elm E, Altman D G, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLOS Med* 2007; 4: e297.
- 11 Bossuyt P M, Reitsma J B, Bruns D E, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Radiol* 2003; 58: 575–580.
- 12 Schulz K F, Altman D G, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Int J Surg* 2011; 9: 672–677.
- 13 Zachariah R, Harries A D, Ishikawa N, et al. Operational research in low-income countries: what, why, and how? *Lancet Infect Dis* 2009; 9: 711–717.

Contexte : Les formations à la saisie de données sont très rares comparées à la formation à l'analyse des données. Nous avons donc voulu combler ce manque en réalisant une formation brève de 2 jours sur le thème « saisie de données de qualité garantie avec EpiData » destinée à des professionnels de santé publique à Bangalore et Pondichéry, en Inde, en 2013.

Objectif : Déterminer la proportion de participants qui ont utilisé EpiData et ont enseigné le logiciel à d'autres pendant les 2 années qui ont suivi la formation et explorer les raisons pour lesquelles ils l'ont fait ou pas fait.

Schéma : Etude transversale grâce à un questionnaire auto-administré, semi-structuré élaboré avec les formulaires Google trouvés sur le web. Nous avons réalisé une analyse manuelle thématique afin d'identifier les raisons majeures d'utiliser ou non EpiData.

Résultats : Sur 46 participants, 38 (83%) ont répondu. Sur 31 participants impliqués dans la recherche, 17 (55%) ont utilisé EpiData, dont 6 (35%) ont réalisé la double entrée et la validation. Sur les 14 participants qui n'ont pas utilisé EpiData, 11 ont utilisé MS Excel ou SPSS/Epi Info pour la saisie des données. Sur 38 participants, 29 (76%) ont enseigné EpiData à leurs collègues et étudiants. Les raisons du choix d'EpiData incluent la convivialité du logiciel, la liberté d'accès et la facilité à éviter les erreurs de saisie de données. À l'inverse, ceux qui ne l'ont pas fait ont invoqué le manque de temps et de main d'œuvre.

Conclusion : Cette formation brève a été utile en termes de transfert de connaissances et fournit un modèle extensible à incorporer dans les programmes d'enseignement des facultés de médecine et des institutions de recherche.

Marco de referencia: Los proyectos de adiestramiento en materia de introducción de datos son infrecuentes en comparación con la capacitación en el análisis de datos. Con el propósito de corregir esta deficiencia se impartió un curso de dos días sobre 'captación de datos con garantía de la calidad con el programa EpiData' dirigido a los profesionales de salud pública en Bangalore y Puducherry, en la India, en el 2013.

Objetivo: Determinar la proporción de participantes que utilizaba el programa EpiData y lo enseñaba a otras personas durante los 2 años que siguieron a la capacitación y analizar las razones de las actitudes observadas.

Método: Fue este un estudio transversal realizado tras la autoadministración de un cuestionario semiestructurado mediante Formularios con Google en línea. Se realizó un análisis temático manual con el fin de reconocer las principales razones para la utilización del programa EpiData o su falta de aplicación.

Resultados: De los 46 participantes, respondieron 38 (83%). De los 31 participantes vinculados con la investigación, 17 usaban EpiData (55%) y de ellos seis practicaban la doble introducción de datos y la validación (35%). De las 14 personas que no utilizaban EpiData, 11 aplicaban MS Excel o módulos SPSS y Epi Info en la introducción de datos. De los 38 participantes, 29 enseñaban el programa EpiData a otros colegas y estudiantes (76%). Entre las razones aducidas para el uso del programa EpiData estaban su sencillez de utilización, el hecho de tratarse de un programa en acceso libre y la facilidad con que se evitan los errores de introducción. Las razones para no aplicar la doble entrada de datos fueron el tiempo y la escasez de personal.

Conclusión: El curso corto fue eficaz en materia de transmisión de los conocimientos y ofrece un modelo escalonable que se puede incorporar en los programas educativos de las facultades de medicina y los establecimientos de investigación.