African Journal of Emergency Medicine 10 (2020) S144–S148

Contents lists available at ScienceDirect

African Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/afjem

Research primer

Research skills and the data spreadsheet: A research primer for low- and middle-income countries

David McD. Taylor[a,b,*], Peter W. Hodkinson[c], Abdus Salam Khan[d], Erin L. Simon[e,f]

[a] Department of Medicine, University of Melbourne, Parkville, Victoria, Australia
[b] Emergency Department, Austin Health, Heidelberg, Victoria, Australia
[c] Division of Emergency Medicine, University of Cape Town, Groote Schuur Hospital, Cape Town, South Africa
[d] Shifa International Hospital, Islamabad, Pakistan
[e] Cleveland Clinic Akron General, Department of Emergency Medicine, Akron, OH, United States of America
[f] Northeast Ohio Medical University, Rootstown, OH, United States of America

ARTICLE INFO

Keywords:
Spreadsheet
Data
Research

ABSTRACT

The specialty of Emergency Medicine continues to expand and mature worldwide. As a relatively new specialty, the body of research that underpins patient management in the emergency department (ED) setting needs to be expanded for optimum patient care. Research in the ED, however, is complicated by a number of issues including limited time and resources, urgency for some therapeutic investigations and interventions, and difficulties in obtaining truly informed patient consent. Notwithstanding these issues, many of the fundamental principles of medical research apply equally to ED research. In all medical disciplines, data needs to be collected, collated and stored for analysis and a data spreadsheet is employed for this purpose. Like other aspects of clinical research, the use of the data spreadsheet needs to be exacting and appropriate.

This research primer explores the choice of available spreadsheets and a range of principles for their best-practice use. It is deliberately, not an exhaustive review of the subject. However, we aim to explore basic principles and some of the most accessible and widely used data spreadsheets.

African relevance

- Clinical research should be most relevant to the population where it is undertaken.
- Research capacity should move forward as a country develops.
- Generation and management of a spreadsheet is a fundamental research skill.

The International Federation for Emergency Medicine global health research primer

This paper forms part 10 of a series of how to papers, commissioned by the International Federation for Emergency Medicine. This research primer explores the choice of available spreadsheets and a range of principles for their best-practice use. It explores basic principles and some of the most accessible and widely used data spreadsheets.

Background

As a relatively new specialty, research in emergency medicine is still developing. Only in the last three decades has substantial research been undertaken specifically in the emergency department (ED) setting. The importance of this lies in that ED patients differ in many ways from those in other settings. They are, by definition, undifferentiated and often affected by anxiety, pain, fear, and vulnerability. Hence, ED research must be undertaken in the ED setting and inferences from other settings are unacceptable.

One fundamental research skill is the generation and management of a data spreadsheet. Essentially, this is an electronic document where data on each enrolled patient or entity under investigation is stored in a systematic way. Spreadsheets can be used for organization, analysis and storage of data in tabular form. They also allow the manipulation of data and the generation of graphics and summary or simple statistics. Where complicated statistical analysis is required, that cannot be done using spreadsheet software, datasets can usually be imported into statistical packages and analyzed further.

* Corresponding author.
E-mail address: David.Taylor@austin.org.au (D.M. Taylor).

https://doi.org/10.1016/j.afjem.2020.05.003
Received 11 December 2019; Received in revised form 23 April 2020; Accepted 6 May 2020
Available online 07 June 2020

Spreadsheets can also serve as data storage facilities. Subsequent access to the data may be required well after its original analysis and publication of the project's findings e.g. secondary data analysis, merger with data from similar projects and the sharing of data with other researchers (an increasing trend) [1,2].

Organizations ensure high quality, ethical research by utilizing research governance [3]. Governance is involved at all stages of research especially at project submission, during patient enrolment and at closure. Requirements also include data storage for many years and it may be audited to ensure good research practice [2].

Which spreadsheet to use

A wide variety of spreadsheets are now available including Microsoft Excel® [4], Libre Office® [5], Open Office® [6] and Google Sheets® [7]. In addition, there are database applications with considerable functionality behind the spreadsheet interface e.g. Access® [4] and RedCap® [8]. As each has its own characteristics and functionality, the researcher needs to consider the ways in which the spreadsheet will be used and this should be an informed decision. It is recommended, however, that the researcher becomes thoroughly familiar with one variety of spreadsheet.

For the novice researcher, Microsoft Excel® [4], Libre Office® [5] or Open Office® [6] will likely more than suffice. Microsoft Excel® [4] is relatively simple, inexpensive, widely accessible and has an increasing range of functionality. A free comprehensive instruction manual is also available on-line [9]. Given this, Microsoft Excel will be used throughout this primer where examples are provided.

Setting up the spreadsheet

Time and care in setting up the spreadsheet prior to data entry is important and should be done in concert with the data collection document. This will facilitate data entry, help to avoid confusion and mistakes and provide the most appropriate formatting for data analysis either within the spreadsheet or when exported to a statistical program. If a biostatistician will be undertaking the data analysis, it is highly recommended that he/she assists in setting up the spreadsheet. This may save considerable time and effort later on. A data collection and management practice instruction document is also recommended to assist in the efficient and accurate data procurement and process [10].

The most commonly used spreadsheet format for studies involving individual patients is to assign each patient to a single row of data cells (Fig. 1). Row 1 is usually reserved for the column headings and patient data in the rows below. Next, the data from patient study identification (ID) number 1 is placed in row 2, data from patient number 2 in row 3 and so on.

Each column is reserved for data on specific characteristics (variables) of the patients (Fig. 1). Commonly, column A is reserved for the patient study ID number and all other data in the columns to the right. It is advisable to group the columns according to the type of data they contain e.g. columns B, C, D and E may contain patient demographics (e.g. age, sex weight, co-morbidities). Other data groups may include:

- other patient characteristics (e.g. ethnicity, religion, employment status)
- study details (e.g. group allocation, date of enrolment)
- potential confounding and bias variables (e.g. triage score, pain scores, language fluency)
- outcome variables (e.g. patient satisfaction, procedural outcome, adverse events)

Within each column, the data may be numerical (e.g. blood pressure), ordinal (e.g. age group) or categorical (e.g. ethnic group). For ease of data analysis, assign ordinal and categorical subgroup variables with a number and use these numbers to populate the cells. For example, in 'laceration depth' column, insert 1 if epidermal, 2 if dermal and so on (Fig. 1). In anticipation of merging data with that of others, unambiguous ('Pat. ID' not 'PTID') or standard variable nomenclature (e.g. 1 = male, 2 = female) should be used.

Determining the number of variable subgroups will depend upon the nature of the study and the relevance of each variable. In general, approximately 5-6 subgroups are appropriate. Where possible, adjust the subgroups to ensure that each contains similar numbers of patients. For example, if the age subgroup of 80–100 years has few patients, that group could be adjusted to 70–100 years. Another approach may be to use actual data in the first data spreadsheet and then form the subgroups. It is advisable to add parameters for the expected data item range in drop down boxes. This will prevent inaccurate data entry e.g. age of 201 entered instead of 101 years.

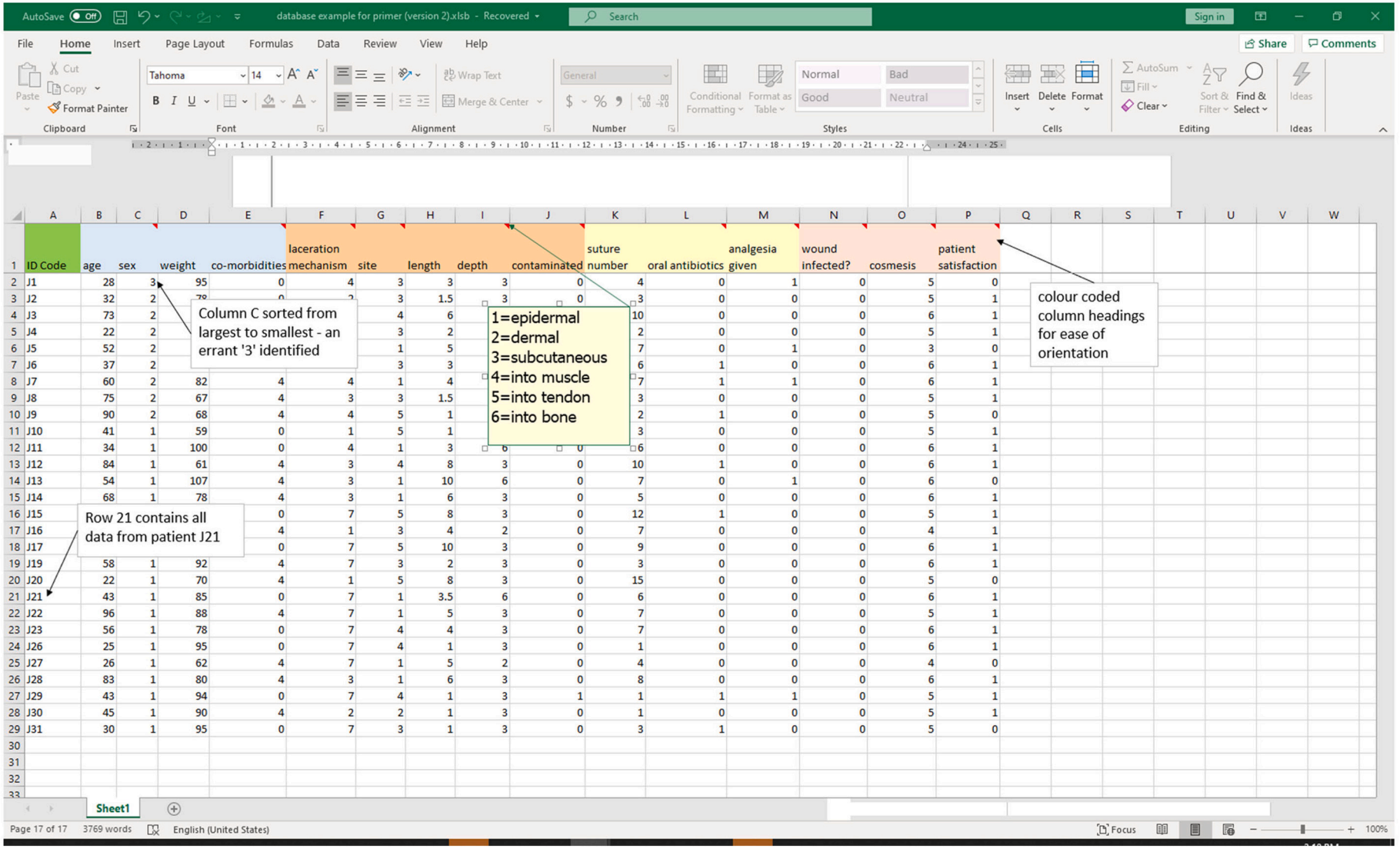
Once the spreadsheet has been set up, it should be trialed before definitive data entry is commenced. This involves entering data for a sample of patients (e.g. 50). At this stage, it is not uncommon to find errors in the way the spreadsheet has been set up or in the way data is to be entered. For example, country subgroups may comprise European, Asian, North American, South American, Australasian and 'other'. However, if the 'other' subgroup is, unexpectedly, found to be large then one or more new subgroups could be added or a new column. Sometimes, there may be two items to be entered into the one cell. For example, pain management subgroups might be oral simple analgesia, parenteral simple analgesia, oral opioid, parenteral opioid etc. However, a patient might have received more than one of these medications. Entering '3,4' in a cell to reflect the nature of the analgesia received, will make data analysis difficult. It is advisable to add additional columns – in this example, the first column will contain '3' and the next will contain '4'.

Navigation around the spreadsheet

Data entry, cleaning and analysis involves moving up and down and across the spreadsheet. While this can become rather confusing, there are tricks to mitigate this. As discussed above, there may be a number of subgroups for each study variable. It is important to devise a way of finding out which subgroup classification a certain cell number represents. This could be on a separate electronic or hard copy form that states 'column F, country of birth: 1 = European, 2 = Asian and so on'. However, this is cumbersome. It is recommended that the column headings have a 'comment' inserted. To do this, left click on the column heading cell (row 1), then right click on the same cell and choose 'insert comment' from the dropdown list. A small dialogue box will appear. Within this box, type the variable subgroups (e.g. 1 = European, 2 = Asian) with each on a separate line. Once done, left click on any other cell. You should then see a small red 'flag' appear on the top right of the cell you have just added the comment to (Fig. 1). If you then move your cursor over that cell (do not click), the dialogue box will appear with all the subgroup details within. You can edit the comment by right clicking on the cell and then choosing 'edit comment'.

Color coding of the column headings (or even the entire column) will help to identify particular data groups. For example, the column headings (in row 1) for the demographic, study detail and outcome variables could be colored yellow, pink and blue, respectively (Fig. 1). This makes finding the column much easier e.g. simply cross to the columns of the appropriate color and search through those ones.

When an Excel® spreadsheet is opened for the first time, moving to the columns on the right may push column A out of view. Similarly, moving down the spreadsheet may push row 1 out of view. When this happens, it is easy to become disorientated if you cannot view the column heading or the patient ID numbers. To avoid this, it is advisable to use the 'freeze panes' function to freeze row 1 and column A in place so they are always visible. To do this, left click in cell B2, then the 'view' tab and then 'freeze panes' in the drop down list. This can be undone at any time and various numbers of columns and rows can be frozen



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID Code	age	sex	weight	co-morbidities	laceration mechanism	site	length	depth	contaminated	suture number	oral antibiotics	analgesia given	wound infected?	cosmesis	patient satisfaction
2	J1	28	3	95	0	4	3	3	3	0	4	0	1	0	5	0
3	J2	32	2	78	0	2	3	1.5	3	0	3	0	0	0	5	1
4	J3	73	2				4	6			10	0	0	0	6	1
5	J4	22	2				3	2			2	0	0	0	5	1
6	J5	52	2				1	5			7	0	1	0	3	0
7	J6	37	2				3	3			6	1	0	0	6	1
8	J7	60	2	82	4	4	1	4			7	1	1	0	6	1
9	J8	75	2	67	4	3	3	1.5			3	0	0	0	5	1
10	J9	90	2	68	4	4	5	1			2	1	0	0	5	0
11	J10	41	1	59	0	1	5	1			3	0	0	0	5	1
12	J11	34	1	100	0	4	1	3	6	0	6	0	0	0	6	1
13	J12	84	1	61	4	3	4	8	3	0	10	1	0	0	6	1
14	J13	54	1	107	4	3	1	10	6	0	7	0	1	0	6	0
15	J14	68	1	78	4	3	1	6	3	0	5	0	0	0	6	1
16	J15				0	7	5	8	3	0	12	1	0	0	5	1
17	J16				4	1	3	4	2	0	7	0	0	0	4	1
18	J17				0	7	5	10	3	0	9	0	0	0	6	1
19	J19	58	1	92	4	7	3	2	3	0	3	0	0	0	6	1
20	J20	22	1	70	4	1	5	8	3	0	15	0	0	0	5	0
21	J21	43	1	85	0	7	1	3.5	6	0	6	0	0	0	6	1
22	J22	96	1	88	4	7	1	5	3	0	7	0	0	0	5	1
23	J23	56	1	78	0	7	4	4	3	0	7	0	0	0	6	1
24	J26	25	1	95	0	7	4	1	3	0	1	0	0	0	6	1
25	J27	26	1	62	4	7	1	5	2	0	4	0	0	0	4	0
26	J28	83	1	80	4	3	1	6	3	0	8	0	0	0	6	1
27	J29	43	1	94	0	7	4	1	3	1	1	1	1	0	5	1
28	J30	45	1	90	4	2	2	1	3	0	1	0	0	0	5	1
29	J31	30	1	95	0	7	3	1	3	0	3	1	0	0	5	0

Fig. 1. Excel spreadsheet of sample data from a laceration repair study.

depending upon which cell you first click on.

Data entry

The transfer of data from its source to the study spreadsheet is commonly associated with mistakes. When done manually, the mistakes usually involve misinterpretation of a data item to be transferred (e.g. entering 456 instead of 654) or typographical errors (e.g. entering 653 instead of 654).

Ideally, data would be transferred electronically e.g. vitals signs transferred directly from ED computers into the spreadsheet [11]. However, this requires sophisticated computer systems and is usually not possible. A reasonable option is to enter data manually, from its source directly into the spreadsheet. In this option, data are extracted from the medical record (or other source) and typed directly into the spread sheet, often via portable devices such as computer tablets or laptops. The least favorable option is manual transfer of data from the source onto hard copy data collection forms and subsequently into the spreadsheet. The more transfer required, the greater the risk of mistakes.

To mitigate the risk of mistakes, data 'double entry' can be employed, where two persons enter the data separately. Upon completion, the two datasets are compared, inconsistencies between the data sets identified, those data are double-checked and reconciled, and the spreadsheet corrected if necessary. However, 'double entry' is resource intensive and this may preclude its use.

An alternative is to scan the data into tables using optical mark recognition (OMR) and optical character recognition (OCR) software. Machine readable forms can be created using this special software. When scanned or faxed, the handwritten information on these forms is read into the database. The advantage is that keyboard entry is eliminated. The disadvantage is that they are more difficult and costly to set up.

In all studies where a single person manually enters the data, a data quality assurance exercise should be undertaken. This involves a second person extracting at least 10% of all data as well (e.g. data from 10% of patients). Like 'double entry', the data from both extractors is compared. Differences (if any) are then reconciled. It may sometimes be necessary to check the entire dataset if more than the very occasional mistake is identified. It should be noted that many journals now require a description of the data quality assurance exercise to be included in the Methods section of the research paper [12]. Failure to undertake and report this exercise will likely result in rejection of the paper.

Data cleaning

Once all data entry is complete and the quality assurance exercise has been undertaken, the data needs to be 'cleaned'. Data cleaning refers to identifying incomplete, inaccurate or irrelevant data and then replacing coarse data with clean entries in a methodical way [13]. In most cases, this involves identifying missing or incorrect data in the spreadsheet. Even if 10% of the data has been checked by a second person, there may be mistakes in the remaining 90% that should be sought.

Rather than scanning every spreadsheet cell to identify inconsistencies, 'range checking' can be undertaken. In Excel®, this technique involves highlighting an entire column of data and clicking on 'sort smallest to largest' (for numerical data) or 'sort A to Z' (for text data) on the 'Sort & Filter' dropdown box on the Home tab. Incorrect data items will be found at either the top or the bottom of the sorted data column. For example, a range check of sex (where 1 = male and 2 = female) may identify an errant '3' (Fig. 1). This value is 'out of range' for the study and needs to be corrected.

Once the range check of a column has been done, 'undo' the sorting before correcting any errors or moving on to the next column. To find the error once the sorting has been undone, highlight the column, click 'Find' from the 'Find and Select' dropdown box on the Home tab. Enter the incorrect value in the 'find what' box (e.g. 13 or 134 in the age example above) and click 'Find Next'. The incorrect cell will be highlighted, the patient's study number determined and the data can be corrected.

If data are collected by several investigators or across different sites, then means and medians should be compared across investigators and sites. If there are substantial differences this can indicate systematic differences in measurement or data collection.

Version control

It is very easy to make mistakes or lose track of progress, especially during the data cleaning process and formatting for data analysis. In this regard, spreadsheet 'version control' is vital with the saving of all versions of the spreadsheet, appropriately named and dated. The importance of this lies in the possibility of mistakes being made in the cleaning or analysis (e.g. forgetting to 'undo' sorted columns, accidental deletion of data). If these mistakes cannot be corrected, then at least it is possible to go back to the immediately preceding version and start again.

It is recommended that the first version of the spreadsheet is saved as its own file before every major data manipulation. For example, once all data has been entered, that file could be named and saved as 'raw data'. To progress with the data, the 'raw data' file should be opened and saved as the next version e.g. 'raw data–cleaned'. Once cleaned and saved, the file is opened and saved as the next version e.g. 'raw data–cleaned–formatted for the statistical software'.

It is recommended that version numbers and dates are also added to the file names e.g. 'v2–raw data–cleaned–11062019'. Having the version number first has the advantage of having all the files stored in the correct version order. This could be lost if the file name comes first, or if it is changed for some reason in a subsequent version. The date is an additional means of tracking the versions if files names are written incorrectly.

Finally, like any electronic documentation, all files need to be backed up in the event of computer failure, theft, fire or other catastrophe. Many facilities have dedicated institutional hard drives on which files can be backed up. An alternative is to use an external computer hard drive. These are now relatively inexpensive, with some having the function of automatically backing up files on a regular basis (e.g. daily). Spreadsheets can also be stored on CD discs or 'in the cloud'.

One consideration for important data spreadsheets (and other files), is to plan for the worst case scenario. Consider the consequences of a fire in your office that destroys your computer, your external hard drive and your storage discs. If these amounted to 3 years of PhD research then it may all be lost. Given such possibilities, it is recommended that important files be kept in at least one remote facility (e.g. your home computer, in the 'cloud').

Confidentiality

Personal data on research subjects must always be treated confidentially. Presently, Institutional Review Boards and Ethics Committees require a description of how the data will be stored confidentially and securely.

In regard to confidentiality, one sound principle is never to have patient identifying information (e.g. name, date of birth) on data collection forms or spreadsheets that also contain their personal data. We recommend setting up a unique identification number ("patient study ID") for each study participant. Using a unique subject identifier that has no meaning external to the study database simplifies the process of "de-linking" study data from personal identifiers for purposes of maintaining subject confidentiality. A separate document called a 'Master List' should be developed. This document will contain and link

the ID numbers with the patients' identities. This may be important if the original patient data source needs to be accessed again to check on data items. The Master List and all other files should be stored separately and should not be shared by all investigators. In the event that either the Master List or the spreadsheet is accessed by an unauthorized person, they will not be able to link patient identity with their data.

In regard to security, all hardcopy data collection sheets should be stored in a locked cabinet within an office that is locked when unattended. Similarly, electronic files, including the study spreadsheets, should be password protected and stored on password protected computers. Only authorized study investigators should be able to access these files.

Tips on this topic

- Prior to data entry, perfect the design, trial and revise the spreadsheet. Invite your statistician to assist at this stage. There are many tutorials on line to assist with spreadsheet set up [14,15]
- Train yourself on the type of spreadsheet that you plan to use. Make mock files and test interactivity, graphs and statistics functions.
- Clean data thoroughly before analyses – this will save time and effort
- Undertake a data quality assurance exercise prior to data analysis. This will ensure high quality data and is required by many journals.

Pitfalls to avoid

- Do not forget the importance of version control and backup of your spreadsheet
- Avoid the possibility of breaches of confidentiality and security of the data
- Avoid multiple persons entering data into the spreadsheet. Also, minimize the number of times data needs to be transferred manually, wherever possible.

Authors' contribution

Authors contributed as follow to the conception or design of the work; the acquisition, analysis, or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content: DT contributed 70%; and PH, ASK and ES 10% each. All authors approved the version to be published and agreed to be accountable for all aspects of the work.

Annotated bibliography

Harvey G. (1) Excel 2016 All-in-one for Dummies. Available at: http://www.allitebooks.in/excel-2016-all-in-one-for-dummies/ (accessed December 11, 2019).

This book is free to download from the Internet. It is a comprehensive guide to the use of Excel. As such it can be somewhat heavy going. It is certainly a reference source when learning new skills but is not for a casual read. A 2019 edition is available.

Declaration of competing interest

The authors declared no conflicts of interest.

References

[1] Gliklich RE, Dreyer NA, Leavy MB. Registries for evaluating patient outcomes: a user's guide [internet]. 3rd ed. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. 6 Data Sources for Registries. Available at: https://www.ncbi.nlm.nih.gov/books/NBK208611/, Accessed date: 31 March 2020.

[2] Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. Shanghai Arch Psychiatry 2014;26:371–5. https://doi.org/10.11919/j.issn.1002-0829.214171.

[3] Slowther A, Boynton P, Shaw S. Research governance: ethical issues. J R Soc Med 2006;99:65–72. https://doi.org/10.1258/jrsm.99.2.65.

4. Microsoft Office Help and Training. Available at https://support.office.com/, Accessed date: 11 December 2019.
5. Libre Office and use of Calc. Available at https://www.libreoffice.org/discover/calc/, Accessed date: 30 March 2020.
6. Apache Software Foundation. Apache Open Office Available at https://www.openoffice.org, Accessed date: 31 March 2020.
7. Google Sheets. Available at https://www.google.com.au/sheets/about/, Accessed date: 11 December 2019.
8. Research Electronic Data Capture (REDCap). Available at https://www.project-redcap.org/, Accessed date: 11 December 2019.
9. Harvey G. (1) Excel 2016 all-in-one for dummies Available at http://www.allitebooks.in/excel-2016-all-in-one-for-dummies/, Accessed date: 11 December 2019.
10. Strasser C, Cook R, Michener W, Budden A. Primer on data management: what you always wanted to know Available at https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf, Accessed date: 10 March 2020.
11. Kalogriopoulos NA, Baran J, Nimunkar AJ, Webster JG. Electronic medical record systems for developing countries: review Available at https://ieeexplore.ieee.org/document/5333561/, Accessed date: 31 March 2020.
12. Annals of Emergency Medicine. Instructions for authors Available at https://www.annemergmed.com/content/instauth, Accessed date: 30 March 2020.

[13] Welch G, von Recklinghausen F, Taenzer A, Savitz L, Weiss L. Data cleaning in the evaluation of a multi-site intervention project. EGEMS (Wash DC) 2017;5:4. https://doi.org/10.5334/egems.196.

14. LifeWire. Spreadsheet tutorials Available at https://www.lifewire.com/learn-how-spreadsheets-4160655, Accessed date: 11 December 2019.
15. wikiHow to do anything. How to make a spreadsheet in Excel: 14 steps (with pictures) Available at https://www.wikihow.com/Make-a-Spreadsheet-in-Excel, Accessed date: 11 December 2019.