How to design and use a research database

Simon S Cross

Ian R Palmer

Timothy J Stephenson

Abstract

The vast majority of histopathology research projects based on a series of tissue samples will require a database to store and organize the data. Most of these databases will be relatively simple standalone databases that can be created in a spreadsheet application. However, there are still some important considerations in the design of the database and coding of the data that will make its subsequent use much more time efficient. This article reviews the scope of databases for common histopathology research projects, the ethical and legal considerations, design of the database structure, coding of data items, and import and export of data from databases.

Keywords anonymization; databases; histopathology

Introduction

Most research projects need some method for recording and ordering data. Sometimes this may be a complex system that automatically links with other data sources, but in most cases it will be a relatively simple standalone database for a single researcher to maintain their research data. However, even though this level of database may be relatively simple, it is still very important to give some thought to its design before starting to use it, as a well-designed database will make analysing results much easier, and a poorly-designed database can lead to much wasted time which may even extend as far as having to re-enter data.

What sort of database does your project need?

At the start of a research project you need to make a simple assessment of what your database will need to do. If your research project is a randomized controlled trial involving several different centres and which requires data to be entered by several different personnel, then we hope you have budgeted for a dedicated database manager in your research grant because you will need them! Such a database is well outside the scope of this paper. A more tractable project might be something like a number of different immunohistochemical stains performed on a retrospective series of 500 cases of breast cancer with histopathological data (such as tumour size, grade, lymph node status) and clinical follow-up (e.g. time to last follow-up, recurrent disease, death not due to breast cancer, death due to breast cancer). A database for a project such as this would be a reasonable proposition for a single researcher, such as a histopathology trainee working on an PhD degree, and is the scale of database that we use all the time in our projects.[1]

Anonymization, ethical and legal considerations

Any research database should be anonymized to the highest level that is compatible with the viability of the research project. The recent high profile cases in the media highlighting data losses by Government agencies has shown how easy it is for data in some format (laptop, hard drive, data stick, and more recently insecure server etc.) to be lost or stolen.[2] It is likely that you will be moving your data around between computers and it is probable that you will be using a laptop computer, so for security considerations alone the database should be anonymized. Any research project will require local ethics committee approval, and most ethics committees are rightly concerned about the transmission of personal medical data and usually insist that any research databases are anonymized.

There are two main methods of anonymization – complete and linked. The most secure is complete anonymization where the clinicopathological data are taken from their original source and entered into the research database with no linkage at all between the two. If all personal identifiers are excluded from the research database, then there should be no way of systematically identifying patients from the data held in the research database (it should, however, be noted that it might be possible to identify occasional individuals if they have very distinctive characteristics, e.g. a 16-year-old female living in city ‘x’ who was diagnosed with breast cancer). The advantage of a totally anonymized database is a lowering of the risk should it be lost or stolen. It obviously reduces the risk of identification of individuals to close to zero, but it does not eliminate the risk of media coverage should the lost data be found somewhere inappropriate (the local newspaper headline ‘Cancer patients’ data found on memory stick in car park’ could still be applied whether the data were anonymized or not). The major disadvantage of a completely anonymized database is that it is a static database that cannot be updated later if there is a need for further data input. This may not be a problem for many histopathology research projects (e.g. novel immunohistochemistry on a series of tumours) but is obviously not compatible with projects with continuing patient follow-up.

In linked anonymization there is some system that links the research database to identifiable patient data. One method of linkage is to have a separate secure database that acts as a key between identifiable data records and the anonymized research database. Such a database does not need to be electronic; it could simply be a book with two entries on each line – the identifiable data and the anonymized research accession number (often known as the study number). Such a book could be kept securely

***Simon S Cross** MD FRCPath is Professor of Diagnostic Histopathology at the University of Sheffield and an Honorary Consultant Histopathologist at Sheffield Teaching Hospitals NHS Foundation, UK. Conflict of interest: There is no conflict of interest.*

***Ian R Palmer** BSc is the Faculty IT Manager (Medicine, Dentistry & Health) and Faculty IT Liaison Corporate Information and Computing Services at the University of Sheffield, Sheffield, UK. Conflict of interest: There is no conflict of interest.*

***Timothy J Stephenson** MA MD MBA FRCPath is a Consultant Histopathologist at Sheffield Teaching Hospitals NHS Foundation Trust and Honorary Professor at the University of Sheffield and Sheffield Hallam University, Sheffield, UK. Conflict of interest: There is no conflict of interest.*

locked in a filing cabinet in a locked office and access could be restricted to a few named individuals involved in the research project; often it will be a single named individual whose sole purpose is to provide linkage if requested. However, this method involves setting up a separate administration system with its own security issues. A much simpler method is to use some identifier which is not identifiable without access to some existing information system. This could be the patient's hospital number but many ethics committees regard this as relatively insecure because the patient's hospital number could open access to a whole range of information in many different hospital information systems (and the patient's NHS number is even less secure and more informative). In histopathology we are fortunate in having the laboratory accession number which is a restrictive superficially-anonymous identifier that is only available to those with access to the laboratory information computer system, and that system is already heavily protected by individual password login, automatic timing out of terminals after a short period of non-use, etc. If we are conducting a research study on archival diagnostic surgical material, then we will probably be using slides labelled with the laboratory accession number and we are unlikely to anonymize that number because of the slide labelling problems it would cause. At present the ethics committees that we have passed applications through have been happy to use the laboratory accession number to anonymize the samples and data in studies, although they have always required a statement on the circumstances in which the investigators would go back to the laboratory information system for more data. Such considerations always have to be made for any research project and are not exclusively a feature of research databases.

It still needs to be appreciated that simple personal data, such as name and address, are not the only items that can identify an individual. The date of birth is quite a specific item and combined with other items in a database (e.g. a woman with breast cancer diagnosed in 2002) could provide a strong indication of an individual's identity. It is usually better to change the date of birth to a less specific age datum at the time of entering information into the database; age at diagnosis will probably give all the information that is required.

It should be noted that even when a database has been 'anonymized' using relatively opaque linkage items, such as the laboratory accession number, it is still not completely anonymized because links can be made indirectly which could identify the patient. Although this linked 'anonymization' will minimize the adverse effects of any data loss or unauthorized access to the data, the database is still legally regarded as personal data under legislation such as the UK Data Protection Act.[3,4] Under this Act, only data essential to the study can be collected and they can only be stored for as long as the study is active.

This brings us to legal consideration as whether the database needs to be registered under the terms of the Data Protection Act in the UK, or similar legislation in other countries. If there is no linkage possible at all between the research database and personal patient data, then it would not need to registered, but since the act of creation of the database will involve some access to personal data, then we would suggest that it does need registration. In most UK institutions this is a simple process because there will be an institutional Data Protection Officer and the institution will be registered with the Government's Information Commissioner, which will usually give generic registration for databases in that institution without the need for the registration of individual databases. However, you would need to check the details of your institution's processes.

A further medicolegal consideration is the duty of disclosure that could occur if review of a case with identifiable data showed that there was a misdiagnosis or mistreatment, e.g. if review of a series of cases of mesothelioma showed that one of the cases appeared to be mesothelial hyperplasia rather than mesothelioma. If there is complete anonymization, then this cannot apply. If there is linked anonymization, then it could apply and this needs to be considered before the study commences. In most histopathology research studies, the tissue being reviewed is only a sample of the whole diagnostic material available on the case, e.g. a single block of breast cancer, so a full diagnostic review of the case is not being carried out. The tissue is also being examined in a research context, e.g. a single core of tumour in a tissue microarray, where it is possible that there could have been some transposition of laboratory accession numbers or rows and columns in a spreadsheet and the risk of misidentification would be slightly higher than in a fully functioning diagnostic histopathology laboratory. For these reasons, we include a statement in our ethics committee application stating that any discoveries are for research purposes only and that there will be no retrograde flow of information back to patient records or individual patients.

Design your database

The design of your database should reflect the design of your research project and it would be better to have all your aims and objectives of the project clearly formulated and recorded before you start to design your database. Sometimes, however, design of the database and general objectives of the research project emerge concurrently. A good way to design your database is to take a blank sheet of A3 paper and sketch out all the data items you think you need in a 'spider' (or 'mindmap') diagram with clusters of related items. This gets you away from a computer screen and a form format, and allows you to look at all the data items with an overview. It is worth spending some time doing this, possibly redrawing the diagram if it becomes too cluttered by changes, because this is the most important stage of database design and time spent on this stage will pay major dividends later in the process.

You need to arrive at a design with all the data items that are essential for your project but no more than that. If you do not include essential data items at the start of the process, you will discover their absence later and then need to modify the database, and you may have to go through all your original data sources to input the missing items. It is usually best to include individual raw data items and use them to calculate derived data items within the database. An example in breast cancer would be the Nottingham Prognostic Index. If you only include the Nottingham Prognostic Index in your database as an aggregate measure of the potential prognostic behaviour of the tumour, then you will not have access to the tumour size, tumour grade and lymph node status in future analyses. If you include those three items separately, then you can create a fourth derived data item that will automatically calculate the Nottingham Prognostic Index from these items. If you only record axillary lymph node status in a case of breast cancer as positive or negative, then you will not have access to the total number of positive lymph nodes, whereas if you record the total number of positive lymph nodes, you will be able to create another automatically derived data item for overall lymph node status. The general rule is that data items cannot be split at a later stage, but they can be easily, and automatically, aggregated later, so it is best to create a database with the simplest raw data items included instead of derivative/aggregated data items.

Any variable that in its raw state is a continuous variable (e.g. tumour size in millimetres) should be stored as that continuous variable rather than a category (e.g. tumours between 5 and 10 mm). You should of course ensure that all the data items in your database are relevant to your project's aims and objectives because it is a waste of resources (usually your time) to enter irrelevant items.

What software should you use?

You will notice that choice of software comes after design of the database. Too many researchers start with the software choice and then their database design is influenced by what the software can do, rather than what the software can do for them. There are a number of popular database software packages on the market, such as Microsoft Access and Filemaker Pro, which run on individual computers. There are also server-based databases, such as MySQL, Microsoft SQL Server or Oracle. These software packages are all very good and very powerful for multiuser projects with a number of interacting databases, but they are overkill for a research project such as the breast cancer tissue array project described above. Although we have used Microsoft Access and Filemaker Pro in the past with success, the learning curve for using these packages is relatively steep and the results are no better than using our current preferred option which is Microsoft Excel. This spreadsheet application has enough database functions built into it to carry out all the necessary preprocessing before statistical analysis of results in specific statistical software package, such as Statistical Package for Social Sciences (SPSS). It has the major advantages of virtually no learning curve, if you already use its spreadsheet functions, no platform specificity (Microsoft Access does not run on Macintosh computers), accepts files from other sources in a wide range of formats, and does not allow the user to waste time creating colourful graphical interfaces for entering data that have no value to the project. We simply create a column for each data item and use a row for each case. The 'Split Window' and 'Freeze Panes' functions (found under the Windows menu at the top of the screen) allow the column headings and left-hand row labels to remain on the screen all the time no matter how many rows or columns there are in the spreadsheet (see the Excel help files for further details). We enter data directly into the spreadsheet without using any custom entry form, but it is easy to create one if you feel that data entry would be cleaner by using one, e.g. if someone without specialist knowledge is entering the data for you.

If you need multiuser entry, then a single spreadsheet becomes problematic because version control becomes very difficult. In these circumstances, a server-based database is more useful but it can be difficult to arrange access from different institutions. A more recent development is the availability of web-based databases, which are accessible from any web browser and make multiuser data entry very simple. The consideration with these web-based databases is the security of the commercial providers' servers, but they are basing their business on a secure service so the risk might be lower than for an ad-hoc server in an academic institution.

Coding your data

You will know which data items you want to include in your database because you designed it before you touched the computer keyboard, but now you need to consider how you are going to code that data. This only requires a bit of thought about how you are going to sort and use the data but again if you code your data in appropriate formats it will cause a lot of rework and delay later. The laboratory accession number is a simple illustrative case (Tables 1 and 2). This number is usually in the form of the year and then a single unique number issued in ascending order as specimens arrive. There may also be a separate block number if the specimen cannot be embedded in a single block. If such a number is entered into a single column in its usual format, e.g. 08/23156/3E, it will be treated as text rather than a number in Excel and there can be a number of problems when sorting data for analysis. Tables 1 and 2 illustrate these problems and show that the best solution is to create separate columns for each element of a multipart or mixed number/text item. Binary data items, such as gender or lymph node status, are easier but there is still a decision to be made as to whether to use strict mathematical binary code or a more descriptive code. Gender could be coded as 1 = female and 0 = male, but this is an arbitrary coding that is not memorable to the data inputer. Coding as f = female and m = male is memorable and so far lower rates of errors are likely during the inputting of the data, i.e. make the purest possible transformations of data. It is a very simple automated 'find and replace' function to convert these to true mathematical binary items for statistical analysis. It is well worth keeping descriptive codings as short as possible, because this reduces the number of keystrokes when entering data. The autofill function of Excel is

Coding the laboratory accession number. If entered in the format in column 1 the accession number sorts to the order in column 2 – the year order is incorrect because the full year was not included. With the full year added and sorted, column 3, there is still a problem with the order of the second and third rows because zeros before the accession number have not been added to numbers with fewer than five digits. When this has been done, column 4, the numbers now do sort into the correct time order but a better method is shown in Table 2

1. Unsorted	2. Sorted	3. Full year sorted	4. Full year and preceeding zeros sorted
98/1119/3E	00/32754/1A	1998/1119/3E	1998/01119/3E
02/34001/4F	02/34001/4F	1999/12237/4D	1999/01231/2F
99/1231/2F	03/4567/8G	1999/1231/2F	1999/12237/4D
00/32754/1A	98/1119/3E	2000/32754/1A	2000/32754/1A
03/4567/8G	99/1231/2F	2002/34001/4F	2002/34001/4F
99/12237/4D	99/12237/4D	2003/4567/8G	2003/04567/8G

Table 1

This shows a better method for coding the laboratory accession number by creating a separate column for each element of the number. Note that in this case preceding zeros do not need to be included as the Excel software is treating the numbers as actual numbers rather than text, as in Table 1. If you are using a single representative section from a multipart specimen, e.g. a representative block of breast cancer from a mastectomy specimen, then always remember to include the specific block number in the database, otherwise future studies may have to retrieve all the slides for the case (specified by the year and accession number) and look at these to select the representative block again — a huge amount of rework for no gain in value

Year	Number	Part
1998	1119	3E
1999	1231	2F
1999	12237	4D
2000	32754	1A
2002	34001	4F
2003	4567	8G

Table 2

useful and can be optimized by choosing different first letters for as many different items as possible. An example for breast cancer that we have used is tumour type, where d = ductal, l = lobular, t = tubular, me = medullary, mu = mucinous, etc. It should be noted that when entering text into Excel it will be case-sensitive in any manipulations, such as grouping or pivot data tables, so it is best to use lower case letter throughout, which also reduces keystrokes by not using the Shift key.

Importing items into your database

If some of the data you wish to have in your database exist in another database, it may be possible to import them into your database instead of having to key them all in. The first thing to ensure is that the data in the other database are 'clean' (i.e. consistently formatted) and verified. People often volunteer data from a database which on closer examinations contains many formatting inconsistencies, e.g. 'yes' coded as 'y', 'Y' or '1' for a single data field, and it then is better to start again with the original source data. Another common feature of processes designed to port data from one database to another is that IT professionals want to write complicated interface programmes which take weeks or months to develop. Whilst these are essential for databases that port data between themselves on a frequent, recurrent basis, this is usually completely unnecessary for a standalone research database, which usually only requires a single transfer at a single point in time. The easiest method we have found is to perform a search on the existing database to select the records that we want to import and then export those in a format that Excel can read. Often the parent database will be able to export in Excel format but comma or tab-delimited text is just as acceptable and causes no extra work. Once in Excel the imported data can be quickly sorted and formatted as required. It should be noted that it is best to import as the first stage of creating your research database, rather than trying to add the imported data to an existing database which may be more problematic.

Calculating derived data items within your database

As noted above it is often easier (and certainly more accurate) to calculate some derived data items from raw source data items within the database. This can be done using Excel's calculation functions and the reader is referred to Excel help files or instructional books for the details of this.[5] Obviously some calculations are easier than others. If your raw data item is the maximum diameter of a breast cancer in millimetres and you want to convert that to centimetres in the calculation of the Nottingham Prognostic Index, then you would simply set up a formula such: = A23/10 where A23 is the index of the cell containing the measurement in millimetres. It is a simple matter to propagate that formula down the whole column for size in centimetres by using the 'Fill' 'Down' functions from the Edit menu. Deriving the lymph node status component for the Nottingham Prognostic Index from the total number of positive lymph nodes would be more complicated but the conditional operators exist in Excel to allow this, e.g. = IF(A23 > 3,2,0), and automatic help boxes make this very easy (click the calculator symbol on the Excel toolbar just to the right of the green tick symbol). It should be noted that these calculated values only exist when the other items from which they are derived are still present in the datasheet. If for any reason, such as reducing data down to the core items for statistical analysis, you are going to remove columns from the datasheet, you must select the whole datasheet, copy it and then paste it back in the same cells using the Paste Special function and selecting Values. You will then have a sheet in which each cell is an independent number not dependent on any other cell. Obviously, you should only do this when you have finished entering all the data and are ready for data analysis, and you should immediately save this sheet in a different name from the original database sheet.

Verifying your data

Once you have input your data you need to make some assessment of the veracity of that data. It is remarkable how many errors are introduced during data entry that need to be corrected before the database is usable. Excel has some useful functions that make debugging the database relatively easy. One method is to ensure that you have saved the database and then try sorting the data in many different ways. Go to 'Sort' under the 'Data' menu and select any single column as the sorting factor and perform the sort. Now scroll through the database sheet looking at the column you sorted by, especially at the top and bottom of the sorted sheet, and you can see all sorts of data entry errors from completely blank cells through to mispellings or duplicated symbols. You could repeat these sorts for each column to check for further errors in other columns. A similar, but possibly easier, method is to use the Autofilter function. Highlight the complete sheet (by clicking the diamond symbol in the top left cell of the sheet), from the Data menu select the Filter option and then select the Autofilter option. At the top of each column there will now appear an up and down arrow symbol; clicking on this produces a list of all the different occurrences of data items in that column, making it very easy to see errors such as 'Y' instead of ‘y’. When all the obvious typing errors have been corrected, it would be wise to perform an audit, checking perhaps 10% of the case records with their original source data to check that there are no systematic data entry errors.

Backing up your database

Data only exists when it is stored in two places — this widespread dictum is never more true than for databases. If you are going to spend many hours entering data into a database, then you need to have a reliable system of back-up. The best systems are automatic and store the data at a geographically separate location; institutions such as universities often provide such systems for their members. Excel-based databases produce files which are very manageable in size and can easily be backed up to CDs or memory sticks, but be mindful of security issues even with a completely anonymized database (which would still cause media interest if found in an inappropriate place) and keep the copies in secure locations. Many NHS institutions are now so wary of data loss through sources such as memory sticks, that they are considering disabling USB ports on their computers so automated back-ups onto servers within the institution’s firewall may well become the norm.

Exporting data for analysis

The function of a database is to store and order data for a research project that can then be analysed. By using Excel as the database software, many statistical functions and analysis can be used directly within Excel. However more sophisticated analyses will require preprocessing and export to other programmes such as SSPS. The preprocessing will mainly consist of converting descriptive variables into numerical values, e.g. ‘f’ to ‘1’, which is easily done using the search and replace functions of Excel. The labels of the columns may have to be shortened to be read by some statistical programmes, e.g. SPSS will only read the first eight characters of variable names. Again, you must remember to save the spreadsheet with a different filename before beginning data preprocessing, as it will destroy some of the data, or certainly its format. Most specialist statistical programmes, such as SPSS, will read Excel files but data can be output in comma- or tab-delimited text from Excel if necessary.

Research notes ‘database’

Any researcher 15 years ago would have had supplementary ‘databases’ for their research projects, usually an index box with cards with the authors and title of published papers and a number that referred to a folder in a filing cabinet. They might also have had a bound book with notes, or possibly a pile of loose papers with various notes — they still might have those on their desk today. A supervised researcher working in a well-organized laboratory should still have a bound book as a record of their experiments and primary data, but there are now more useful ways of storing other information than a pile of loose papers. The widespread availability of electronic PDF versions of published paper and the information displayed on numerous authoritative websites has widened the scope for electronic databases of collected research ‘notes’. These should be much easier to search, can have direct links to PDF files stored on the computer and direct links to relevant websites, and could significantly improve a researcher’s productivity. A number of software programmes that perform this task are available. On Macintosh computers TapForms[6] appears to be the current market leader and is extremely easy to use. On PCs Microsoft Office One-Note (available as part of the MS Office 2016 Office Suite) is a comprehensive electronic documentation application that allows organization and storage of information from a wide range of sources.[7] It also allows resources to be shared securely with project collaborators via the internet.

Conclusions

A research database for even relatively sophisticated histopathology research projects, e.g. immunohistochemical staining of tissue arrays with hundreds of tumours, is easy to construct and use provided that care is taken in its overall design and the coding of data items. The purpose of the database should always be to serve the overall aims of the research project and it should not become an isolated resource-consuming entity. For most histopathology projects, a simple spreadsheet programme, such as Microsoft Excel, will provide all the necessary database functions with the minimum amount of infra-structural effort.

Conflicts of interest

The authors have no conflicts of interest to declare. ◆

REFERENCES

1 Balasubramanian SP, Cross SS, Globe J, Cox A, Brown NJ, Reed MW. Endostatin gene variation and protein levels in breast cancer susceptibility and severity. *BMC Cancer* 2007; **7:** 107.

2 BBC News website. Thousands of Welsh NHS staff’s data stolen in hack. Avaliable at: http://www.bbc.co.uk/news/uk-wales-39249975 (accessed 4 June 2017).

3 Data Protection Act. Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents (accessed 4 June 2017).

4 Iversen A, Liddell K, Fear N, Hotopf M, Wessely S. Consent, confidentiality and the data protection act. *BMJ* 2006; **332:** 165e9.

5 Walkenbach J. Excel 2016 bible. Hoboken, New Jersey: John Wiley & Sons, 2015.

6 TapForms. Avaliable at: https://www.tapforms.com/ (accessed 4 June 2017).

7 Microsoft Office OneNote. Avaliable at: https://www.onenote.com/ (accessed 4 June 2017).

Practice points

- Any research database should be anonymized as completely as is feasible for the specific research project
- A research database should always be part of a project approved by a research ethics committee and should be specified within the research protocol
- A research database should comply with appropriate data protection legislation, the Data Protection Act in the UK
- The design of the database should be specified by the aims and objectives of the research project and not by software capabilities
- Care should be taken to code data items in a format that retains all potentially relevant information and which allows them to be selected and sorted in a consistent manner