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.*

Please cite this article in press as: Cross SS, et al., How to design and use a research database, Diagnostic Histopathology (2017), https://doi.org/10.1016/j.mpdhp.2017.09.005

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

Please cite this article in press as: Cross SS, et al., How to design and use a research database, Diagnostic Histopathology (2017), https://doi.org/10.1016/j.mpdhp.2017.09.005