Author/s:
Dipnall, JF;Berk, M;Jacka, FN;Williams, LJ;Dodd, S;Pasco, JA

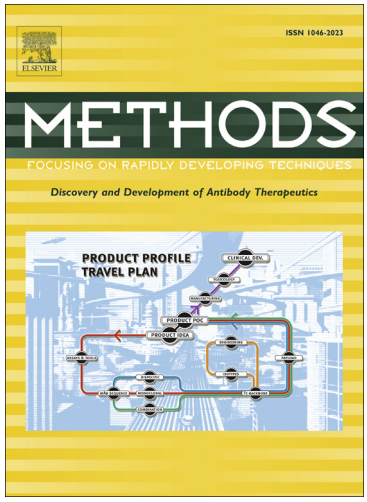
Title:
Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers

Date:
2014-10-01

Citation:
Dipnall, J. F., Berk, M., Jacka, F. N., Williams, L. J., Dodd, S. & Pasco, J. A. (2014). Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers. Methods, 69 (3), pp.237-246. https://doi.org/10.1016/j.ymeth.2014.07.001.

Persistent Link:
https://hdl.handle.net/11343/43863

Accepted Manuscript



Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers

Joanna F Dipnall, Michael Berk, Felice N Jacka, Lana J Williams, Seetal Dodd, Julie A Pasco

PII: S1046-2023(14)00238-2
DOI: http://dx.doi.org/10.1016/j.ymeth.2014.07.001
Reference: YMETH 3460

To appear in: *Methods*

Received Date: 10 February 2014
Revised Date: 2 June 2014
Accepted Date: 5 July 2014

Please cite this article as: J.F. Dipnall, M. Berk, F.N. Jacka, L.J. Williams, S. Dodd, J.A. Pasco, Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers, *Methods* (2014), doi: http://dx.doi.org/10.1016/j.ymeth.2014.07.001

Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers.

Joanna F Dipnall[1,2], Michael Berk[1,3,4,5], Felice N Jacka[1,3], Lana J Williams[1,3], Seetal Dodd[1,3], Julie A Pasco[1,6]

jdipnall@deakin.edu.au

mikebe@barwonhealth.org.au

felice@BarwonHealth.org.au

lanaw@barwonhealth.org.au

seetald@barwonhealth.org.au

juliep@barwonhealth.org.au

[1]IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Victoria, Australia

[2]Faculty of Life and Social Sciences, Swinburne University of Technology, Hawthorn, Victoria, Australia

[3]Department of Psychiatry, The University of Melbourne, Parkville, Victoria, Australia

[4]Florey Institute of Neuroscience and Mental Health, Parkville, Victoria, Australia

[5]Orygen Youth Health Research Centre, Parkville, Victoria, Australia

[6]NorthWest Academic Centre, Department of Medicine, The University of Melbourne, Parkville, Victoria, Australia

ABSTRACT

Introduction: The exponential increase in data, computing power and the availability of readily accessible analytical software has allowed organisations around the world to leverage the benefits of integrating multiple heterogeneous data files for enterprise-level planning and decision making. Benefits from effective data integration to the health and medical research community include more trustworthy research, higher service quality, improved personnel efficiency, reduction of redundant tasks, facilitation of auditing and more timely, relevant and specific information. The costs of poor quality processes elevate the risk of erroneous outcomes, an erosion of confidence in the data and the organisations using these data. To date there are no documented set of standards for best practice integration of heterogeneous data files for research purposes. Therefore, the aim of this paper is to describe a set of clear protocol for data file integration (Data Integration Protocol In Ten-steps; DIPIT) translational to any field of research.

Methods and Results: The DIPIT approach consists of a set of 10 systematic methodological steps to ensure the final data are appropriate for the analysis to meet the research objectives, legal and ethical requirements are met, and that data definitions are clear, concise, and comprehensive. This protocol is neither file specific nor software dependent, but aims to be transportable to any data-merging situation to minimise redundancy and error and translational to any field of research. DIPIT aims to generate a master data file that is of the optimal integrity to serve as the basis for research analysis.

Conclusion: With linking of heterogeneous data files becoming increasingly common across all fields of medicine, DIPIT provides a systematic approach to a potentially complex task of integrating a large number of files and variables. The DIPIT protocol will ensure the final integrated data is consistent and of high integrity for the research requirements, useful for practical application across all fields of medical research.

Keywords: data integration, data linkage, merging, data aggregation, data mining, standard.

1. INTRODUCTION

The exponential increase in available data, computing power and the availability of readily accessible analytical software has allowed organisations around the world to leverage the benefits of integrating multiple heterogeneous data files for enterprise-level planning and decision making [1]. The growth of data analytics [2] has meant that organisational information flows have become more targeted and focussed. Benefits from effective data integration include more trustworthy research, higher service quality, improved personnel efficiency, reduction of redundant tasks, facilitation of auditing and more timely, relevant and specific information. Considerable resources are being invested in quality initiatives surrounding data integration; however, poor quality processes underpinning these analytics elevate the risk of erroneous outcomes. The result can be wasted resources and, ultimately, an erosion of confidence in the data and the organisations using these data. The sharing of information can potentially improve policy-making and integrated public services [1, 3].

Data file integration has enhanced knowledge across a broad spectrum of health and medical research, such as health employee research [4], behavioural survey data [5], social sciences [2], patient hospital records [4, 6-8], cancer and other health research [9, 10] and in bio molecular systems [11-14], genetics and genomics [15, 16]. In the field of health and medical science it is becoming an increasing requirement to integrate or merge extensive heterogeneous data files for research purposes, and data files can be linked from multiple providers to perform complex analyses [17-19]. For example, patient data are compiled from both institutional and community settings, including patient records, digital scans, observational surveys, behavioural surveys and official records, and these are often available in diverse and fragmented formats [6].

The plethora of analytical functions required to effectively and accurately integrate heterogeneous data files is challenging and sometimes overwhelming. Often significant funds are invested in quality initiatives that rely on data integration, but variable methodology and thus quality underpinning these analytics elevates the risk of erroneous outcomes. However, to date there are no documented set of standards for best practice integration of heterogeneous data files for research purposes. Therefore, the

aim of this paper is to describe a set of clear operational protocol for data file integration (Data Integration Protocol In Ten-steps; DIPIT).

1.1. The Practice of Data File Integration

Even though the concept of integrating many files to form a single data file for analysis [20] appears relatively straightforward , the actual integration requires careful preparation and a systematic approach to ensure the resulting data are in the correct format, appropriate for the analytical task.

The management process involved in producing a reliable and robust integrated data set from multiple sources, with varying heterogeneous formats, is fraught with potential traps. Large organisations, dedicated to providing data file integration services, have proliferated over the last few decades. The potential to enrich knowledge rises as data integration complexity increases, and with this, potential pitfalls increase. As data files expand in volume and complexity, problems can compound to negatively influence the quality of the final integrated data. The requirements for careful management of the merging and organisational processes are often underestimated, but imperative for reliable results. Many data files often exhibit considerable noise or meaningless data, missing information and unstructured text. All these problems need to be addressed when integrating data [6].

2. METHODS AND RESULTS

The DIPIT approach consists of a set of systematic methodological steps (Table 1) to ensure that: the final data are appropriate for the analysis to meet the research objectives; legal and ethical requirements are met; and that data definitions are clear, concise, and comprehensive. This protocol is neither file specific nor software dependent, but aims to be transportable to any data-merging situation to minimise redundancy and error. It aims to facilitate the generation of a master file that is of the optimal integrity to serve as the basis for analysis.

2.1. DIPIT Step 1: Define the data requirements

It is fundamental to define an hypothesis [21] as data are compiled appropriately and/or analysed in order to test these hypotheses. For example, hypothesis tests determine if a novel treatment is efficacious compared to a control treatment [22]. Costs, in terms of outcomes for individuals, plus

time, dollars and credibility for governments and organisations, can be high if the findings are founded on bad data. Thus, part of the initial step in the assessment of the research data requirements is to establish what data are needed and evaluate the quality requirements for the analytical task, as the final quality of the data may influence the outcomes.

Poor quality and fragmented data are often a result of the compilation of a combination of both manual paper transcribing and electronic entry (i.e. mixed mode). In medicine for example, both paper and electronic medical records are still used in some organisations, which results in fragmented data information [23]. On the other hand, good quality data positively correlates with its use - the better it is, the more it will be used; "data quality and data use are interrelated" [24]. This underscores the importance of assessing the quality of the data to be integrated, and the alignment of the available data to the research objectives, before commencement of the integration process.

2.2. DIPIT Step 2: Establish ethical, legal and privacy issues

Once the file requirements are established, it is important to ensure that legal, ethical and privacy issues are understood and met. For example, the management of personal records are governed by a broad array of guidelines, national and international, specific to different areas and with different levels of authority [25-30]. Neutal [26] highlights potential civil and criminal penalties for violating an individual's privacy, as the public's perception of integrating their own data with many sources is seen as potentially "high risk". Ensuring that the data files used in the integration process comply with standards relevant to the data source is also very important to the public confidence [31]. The imperative for responsible research, sensitive to the rising public awareness of an individual's rights to privacy, has resulted in the establishment of local and international research standards. In Australia, *The Australian Code for the Responsible Conduct of Research* by the National Health and Medical Research Council (NHMRC) provides for responsible research practices and promotes research integrity [32]. This code endeavours to promote integrity in research and define community expectations. The NHMRC code stipulates that clear and accurate records of the research methods, approvals, grants, and data sources during and after the research process be kept, indexed and easily

retrieved if needed. Therefore, it is imperative that data management be performed and documented in accordance to the ethical protocols and relevant legislation related to each file.

2.3. DIPIT Step 3: Order the files to integrate

After establishing the nature and legitimacy of the required files, establish a flowchart for the order of the files to integrate. The flowchart serves as a fundamental document from which the other DIPIT steps feed to catalogue or document the integrated data.

Flowcharts are an invaluable tool regularly used in the fields of Information Technology (IT), Biology, Chemistry and other disciplines for explaining complex mechanisms [33]. Flowcharts simplify complex tasks, and improve comprehension and accuracy [34, 35]. Crews [36] studied the effect that flowcharts had on novice computer programmers, finding the incorporation of a flowchart reduces error and project time.

Since flowcharts are an effective tool for clearly explaining complex models in a comprehensible manner, they lend themselves easily to the documentation of the integration of complex heterogeneous data files and the associated syntaxes. Many statistical computer programs require some form of programming syntax for complex merging requirements (e.g. SAS [37], Stata [38]) and a flowchart of the integration process will develop the appropriate processes. Often computer syntax is used to create the final file efficiently and accurately, ensuring reproducibility (e.g. Stata *do* files, SPSS *sps* files). For example, suppose a government health organisation conducted a study of a specific cohort of patients in hospitals across the nation. Hospitals collected three sets of separate information on each patient. There are two resulting types of data to analyse:

- Data from patient scans consisting of two data files <*ScanA-Filename*> and <*ScanB-Filename*>, and
- Behavioural data from a national postal survey <*SurveyA-Filename*> of hospital patients, and patient structured clinical interviews conducted by hospital doctors <*SurveyB-Filename*>

The objective is to integrate these four files into one *master* data file called <*Master-Filename*> for analysis and a flowchart is established to represent the files to integrate (example shown in Figure 1).

2.4. DIPIT Step 4: Establish the file formats

Once the files to integrate have been established, the formats of each file are identified. Data file formats range from the simple symmetrical text format, comma separated variable format or matrices, relational databases, to a more complex proprietary program-based format (e.g. Stata, SAS) and digital imaging (e.g. DICOM) [37-40]. Some computer programs allow the user to integrate in different file formats [41, 42], but it is often simpler to manage and debug if all the files are converted to the same format. However, many computer applications require input data from multiple sources be in a specific format before the data can be integrated [41, 43].

In the previous example, suppose each of the four source data files had varying formats:

- The two medical scan files have comma separated variable (CSV) and fixed ascii text format (TXT)
- The other two survey files is in CSV format and SPSS proprietary statistical (SPSS) file format

The objective is to integrate the four files into one *Master* integrated file of the proprietary statistical format for analysis. This statistical program merge function only merges using its proprietary data file format, so the four hypothetical heterogeneous data files will be converted to the proprietary file format. The final stage will integrate these preliminary data files into the *Master* data file.

2.5. DIPIT Step 5: Define the variables of interest

Large data files often contain an array of variables of differing importance and label formats. In large data mining exercises there are potentially thousands of variables available for data mining. Avoid including extraneous data that would result in an unnecessarily large *Master* data file. Establishing which variables are of interest and discarding the rest speeds up processing and minimises errors. Eliminating unnecessary variables make the final data file more efficient to search, manage and analyse [44]. So, once the file formats are compiled and documented in the flowchart, the next step is to identify and name the variables of interest.

As with paper records, it is often best if electronic records are well-organised and labelled correctly to ensure ease of identification and accountability. Accurate file naming aligns with the efficient management of the integration of electronic files [44]. Define a global naming convention that easily identifies the source of each variable as this will ease tracking and auditing. As with other electronic libraries, an essential aspect of a well-organised strategy for this step is to standardize filenames used in the integration process. [45].

The variables of interest should be stored and named according to the pre-defined naming convention. Accountability via documentation is an important aspect of DIPIT. A table of variable translations for future reference is established containing at least four columns (example Table 2):

- Final variable name to be stored in the *Master* file and used in the analysis
- Original variable name so that it is always possible to track the source of the variable
- Source file of variable to identify where the variable was originally merged from
- Preliminary file(s) for variable (if applicable)
- Description of variable, including units of measurement

2.6. DIPIT Step 6: Set up link(s) for integration

Effectively merging data files requires the files involved to have one or more common links, whilst ensuring the privacy of a respondent. Linking files with minimal inaccuracy or error is a challenge [29, 46-48]. When multiple merge files have the same single unique identifier, the linkage process is straightforward. Complexity arises when there is no single unique identifier and/or missing data associated with link(s). Data files can be purposefully or accidentally corrupted and/or heterogeneous in nature, causing matching to become an empirical, statistical and often complex challenge [46].

In some instances, integration of data records from different areas, sometime ubiquitous in nature, is a necessary requirement. The National Institute of Standards in the United States of America has recognised the importance of establishing a Unique Patient Index (UPI) and funded research to accelerate the development of a “massively distributed” Master Patient Index (MPI), equivalent to a UPI [29]. However, the ability to link and integrate one or more heterogeneous data files for research

purposes can be hampered by the lack of a common identifier. String variables, such as name, time and date can be difficult to collect correctly and often ambiguous by definition. Inconsistencies in details would deem that data as an unreliable as a link variable (e.g. incorrect date, transcription error) [49, 50]. In health research, linking a patient's clinical characteristics to service utilization can be challenging [6]. The lack of standardization, degree of incomplete information and/or contradictory nature of service information can distort the analysis.

It is important to document the links used for integration for each process and record whether a manual or an automated method was used. Several automated methods have been developed over recent decades to deal with file linkage, such as frequency ratios [51], stepwise deterministic linkage strategies (SDS) [19, 52], probabilistic maximum likelihood linkage models [18, 53], Expectation-Maximization (EM) algorithms[48] and distance-based metrics [54]. It is safe to assert that as the number of files to integrate increases, the linkage challenge compounds and the potential for missing data rises. Weiner et al [17] proposed a "practical, deterministic method of linking Medicare claims" to a large United States clinical database. In the absence of the main United States Social Security Number (SSN) unique patient identifier, an algorithm was used that incorporated a combination of patient attributes (e.g. sex, date of birth, elements of name) to link the data files. In the late 1980s and early 1990s the United States Bureau of the Census developed a record-linkage methodology and automated software system [48] using a linear sum assignment model.

2.7. DIPIT Step 7: Document the integration path

Once the file linkage definitions have been established, the existing flowchart should be expanded to incorporate the systematic hierarchy of the integration process. Most computer programs allow for user-written syntax to manage and automate the analytical process [37, 38, 55, 56]. Designing a logical systematic set of syntax files for merging ensures that all required merging steps will be undertaken and the results will be reproducible.

There are many benefits of using syntax files [37, 38]. The syntax file is an automatic documentation of the integration process that increases efficiency and productivity, while ensuring reproducibility of results.

As a minimum the structure of the syntax file should include:

- The integration of the primary files
- Saving the Master file format in a standard file naming structure
- The variables of interest to be retained
- The variables standard naming format
- Merging of all files into the Master file
- A valuable log of statistics of the key variables, and missing data analysis

An example of a flowchart of the integration process is presented in Figure 2.

2.8. DIPIT Step 8: Flowchart the type of integration

Even though there are a multitude of ways data files can be related or matched [57], the consensus is that there are fundamentally four integration scenarios, characterized by how observations relate to each other amongst the data sets [37, 38, 58]. A simple example of each of these scenarios is shown in Figure 3 based on a small number of file records.

These scenarios assume that there is common information amongst the multiple sources of input data, either at the physical or logical level. Often the *Merge* and *Master* data files do not have the same number of variables. Once integrated, the variables not present in the data files should be coded as missing in the combined master data set. The common integration scenarios are as follows:

- *One-to-One Matching*: This is the most straightforward of the matching scenarios where a single observation in one data set is related to a single observation from another, based on the values of one or more selected variables. For example, in Figure 3, *One-to-one based on one variable* the ID001, ID002 and ID005 are a perfect match. Those IDs that do not match create missing data in the Master file (e.g. ID003, ID006). However, many integration tasks do not have this straightforward matching. *One-to-one* based on two variables, ID001 and ID002 match perfectly, but those IDs that do not match again create missing data in the Master file (e.g. ID001 ACD, ID003 ABC).

- *One-to-Many and Many-to-One Matching:* This implies that one data set has at most one observation with a specific variable or combination of variables, but the other input data set can have duplicates of each value or combination of values. However, the opposite scenario is where there is a many-to-one matching situation. The merge file contained multiple occurrences contained and the master file contains only a single record of each unique identifier, hence the many-to-one scenario.
- *Many-to-Many Matching*: By default the nature of this type of merging implies that multiple observations from each merge file are related based on values of one or more common variables. The developers of the Stata statistical software state: "is difficult to imagine an example of when it would be useful" and do not recommend this type of merging [38] as there are no variables that uniquely identify the observations in either the merge or master data files. However, many-to-many matching still considered a valid form of matching by in other systems, such as SAS [37, 58] and SQL. Moreover,, this form of matching has been utilised in healthcare insurance, where insurance claims are merged with an enrolment database in order to identify the eligibility of the service using the member number as the unique link. Since a member potentially has multiple claims during a certain time period and also many enrolment records, many-to-many match is used in this instance [59].

2.9. DIPIT Step 9: Document the integration outcome

Matching records is simple in theory, but can be complex and tedious in practice. The complexity of mismatching records multiplies when more files and/or merge keys increase, and possibly a mix of scenarios (i.e. one-to-many, many-to-one). A systematic and traceable approach is required so that at any point the origins of the final master data file can be traced and explained.

Assessing the quality of any linkage algorithm used entails determining how many truly matched and non-matched records have resulted. One valid method for evaluating record mismatches in one where each record has a merge identifier indicating its merge status one for integration process. There are many quality measures that can be used, such as accuracy, precision, recall, F-measure graphs and false positive rate [60]. For example, Xu et al utilised a probalistic data linkage method to link data

files in order to investigate major depressive disorders in the perinatal period; in this case they utilised the false positive and false negative rates from a random sample of 1,000 [61, 62].

2.10. DIPIT Step 10: Check variables and missing data

Once the final master data file has been created, an initial inspection of the data is necessary to ensure the integration has been successful, quantify the degree of missing data and to develop the best way forward for analysis. An initial inspection of the final variables integrated, to check such indices as bases, ranges and distributions, is mandatory [63]. Missing data is a common problem with data file integration and affects the statistical analysis performed (e.g. reduces statistical power, increases type I error, introduces bias) [64]. Woods et all identified 71 randomized controlled clinical trials, of which 89% reported having some form of missing outcome data, and emphasised the importance of reporting both the amount and handling of the missing outcomes [65]. Thus, this final step involves performing basic inspection of the key variables and assessing the impact of missing data generated from the merging process(s). This important inspection can be both analytical and graphical, as the reporting of research studies should encompass the details and percentage of missingness [65-67].

Should the integrated data files be mutually exclusive, it is recommended as part of the merging process that the amount and nature of missing data in each of the data files be identified and documented. Often the final integrated master file is heterogeneous, extensive and complex. Therefore, initially identifying the degree of missing data in stages has the advantages of:

- Providing a methodical process of missing data identification and documentation
- Simplifying each task by breaking the task into small steps
- Minimising the potential for error

Should the data files not be mutually exclusive, then it is recommended to perform the integration first then review the missing data. This would be relevant if the handling of the missing data is dependent on a multiple set of variables across the integrated data files.

Missing data are “ubiquitous to all clinical studies” [64] and this fact raises the issue of how to accommodate for the inadequacy of any given data set (e.g. imputations, mixed-effects regression

models [68-70]). Reduced statistical power and potentially biased measures (e.g. parameter estimates, measures of central tendency), caused by losing data due to missing cases, can seriously affect the integrity of the study [71-73]. In medicine, gene research [74], self-reported medical scales [75] and longitudinal clinical trials [69] have experienced issues with missing data when running analyses.

There are many methods of handling missing data, such as i) deleting observations with missing data on any variable (i.e. listwise deletion); ii) deleting observations with missing data on only the two variables of interest (i.e. pairwise deletion); iii) substituting the missing value with the mean of the values of that variable (i.e. mean substitution); iv) substituting the missing value with a predicted value from regression (i.e. regression imputation); v) substituting the missing value with the expected value based on Maximum Likelihood (ML) estimation; and vi) performing a simulation-based procedure in order to handle missing data in a way resulting in valid statistical inference (i.e. ML) [38]. The choice of method used to deal with missing data can influence the analysis performed (e.g. the size and direction of the correlation coefficient). However, Tabachnick and Fidell believe the choice of method used for dealing with missing data less important when the proportion of missing data is less than 5% [63]; however, as the proportion of missing rises, then the choice of technique becomes more important.

There have been a number of missing data procedures suggested in the literature over the last several decades [71, 76-79]. The multiple imputation framework of inference for missing data was developed by Rubin [78] in application to survey nonresponse and Schafer [69] ML estimation is sometimes used to treat missing data as random variables taken out of the likelihood function as if not sampled.

It is also possible to graphically represent missing data to review. Various graphical tools can be used to further understand the nature and degree of missing data. Figure 4 presents graphical representations of missing data for a particular set of 58 variables. From this image the degree of complete and missing data across all variables can be identified: 73 complete with no missing variables and 22 with all 58 variables missing. Alternatively, there is also an example of using data mining graphical techniques to scan the complete integrated file and produce a surface plot of the missing data. The white areas represent missing data and there is a clear missing data pattern across

approximately one third of this set of data requiring further analysis. These graphical tools identify if the missing data is sympomatic of the integration process (e.g. a destinct *unnatural* pattern) or due to the underlying nature of the study (e.g. gender filters in a survey).

3. Example use of DIPIT

DIPIT was used to integrate data files from the National Health and Nutrition Examination Surveys (NHANES), a United States population-based cross-sectional study, with the research objective to study selected demographic, examination and laboratory risk factors for depression. Table 3 outlines the tools used at each DIPIT step for the integration of the selected 80 demographic, examination and laboratory data files downloaded from the NHANES website based on the guidelines provided [80]. Primary NHANES data files were in SAS format and converted into Stata format for the integration process and the final Stata master file used for future statistical analysis. Microsoft Excel and PowerPoint were used for the DIPIT documentation and flowchart purposes.

All DIPIT steps were followed for the data file integration process. This example demonstrated that using DIPIT:

- Legal and ethical requirements were met
- A systematic and comprehensive approach was followed
- Processes of data file integration were documented and thus reproducible for scientific rigor
- The origin of all variables in the master file were documented to ensure traceability
- Missing variable data was quantified for future research analysis
- Data definitions will be clear, concise, and comprehensive
- A transparent, accountable data trail resulted

4. CONCLUSION

The integration of 80 selected demographic, examination and laboratory data files, downloaded from the NHANES website, has been used to highlight how DIPIT ensures that the final data file was appropriate for the research objectives, that the legal and ethical requirements are met, that data

definitions are clear, concise, and comprehensive, and that linkage quality and missing information was identified and addressed.

The linking of as set of heterogeneous files is becoming increasingly common across all fields of medicine. This paper presents a protocol, called DIPIT, to provide a systematic approach to a potentially complex task of integrating a large number of files and variables. Ten steps are proposed in a table format. At each step tools such as tables, flowcharts and log files are required to ensure the integration process is clearly documented and easily reproduced, affording a useful protocol for the accurate and efficient integration of large datasets.

List of abbreviations

DIPIT – Data Integration Protocol In Ten-steps

CSV – Comma Separated Variable

EM – Expectation-Maximization

ML – Maximum Likelihood

MPI – Master Patient Index

NHMRC – National Health and Medical Research Council

SDS – Stepwise Deterministic linkage Strategies

SPSS – SPSS proprietary statistical software

SSN – Social Security Number

TXT – Fixed ascii text format

UPI – Unique Patient Index

Competing interests

There are no competing interests with regards to this manuscript.

Authors' contributions

JFD conceived and designed DIPIT and drafted the manuscript. MB, FNJ, LJW, SD and JAP critically appraised the manuscript. All authors read and approved the final manuscript.

Authors' Information

JFD is a PhD student with the School of Medicine at Deakin University and sessional academic, lecturing in statistics, with the Faculty of Life and Social Sciences at Swinburne University of Technology.

MB is currently a NHMRC Senior Principal research Fellow, and is Alfred Deakin Chair of Psychiatry at Deakin University, where he heads the IMPACT Strategic Research Centre. He also is

an Honorary Professorial Research fellow in the Department of Psychiatry, the Florey Institute for Neuroscience and Mental Health and Orygen Youth Health at Melbourne University, as well as in the School of Public Health and Preventive Medicine at Monash University.

FNJ is a psychiatric epidemiologist with a research focus on the prevention of mental disorders

LJW is currently an Alfred Deakin Postdoctoral Research Fellow and Head of the Psychiatric Epidemiology group within the IMPACT Strategic Research Centre, School of Medicine at Deakin University.

SD is a Clinical Associate Professor with the School of Medicine at Deakin University and a Senior Fellow with the Department of Psychiatry at the University of Melbourne. He is part of IMPACT Strategic Research Centre where he is involved with numerous mental health research project. His research interests include; pharmacotherapy, drug safety and the biological underpinnings of mental disorders.

JP is an epidemiologist who investigates healthy ageing and the role of body composition in the development of chronic disease and the role of inflammation in linking comorbid physical and mental ill-health

References

[1] J. Zeng, Center for Technology in Government University at Albany/SUNY, (1999).
[2] D. Bollier, C.M. Firestone, The promise and peril of big data, Aspen Institute, Communications and Society Program Washington, DC, USA2010.
[3] S.S. Dawes, J. Policy Anal. Manage., 15 (1996) 377-394.
[4] P.E. Spector, S.M. Jex, J. Appl. Psychol., 76 (1991) 46.
[5] S.L. Puller, L.A. Greening, Energy Economics, 21 (1999) 37-52.
[6] A.B. Rothbard, A.P. Schinnar, T.R. Hadley, J. Rovi, Administration and Policy in Mental Health and Mental Health Services Research, 18 (1990) 91-99.
[7] P.E. Spector, D.J. Dwyer, S.M. Jex, J. Appl. Psychol., 73 (1988) 11.
[8] Y. Kiyota, S. Schneeweiss, R.J. Glynn, C.C. Cannuscio, J. Avorn, D.H. Solomon, Am. Heart J., 148 (2004) 99-104.
[9] A. Daemen, O. Gevaert, T. De Bie, A. Debucquoy, J.-P. Machiels, B. De Moor, K. Haustermans, Pac. Symp. Biocomput.2008, pp. 166-177.
[10] J.K. Choi, U. Yu, S. Kim, O.J. Yoo, Bioinformatics, 19 (2003) i84-i90.
[11] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, Nat. Biotechnol., 25 (2007) 1251-1255.
[12] S.P. Akula, R.N. Miriyala, H. Thota, A.A. Rao, S. Gedela, Bioinformation, 3 (2009) 284.
[13] C.F. Quo, C. Kaddi, J.H. Phan, A. Zollanvari, M. Xu, M.D. Wang, G. Alterovitz, Briefings in bioinformatics, 13 (2012) 430-445.
[14] J. A Seoane, V. Aguiar-Pulido, C. R Munteanu, D. Rivero, J. R Rabunal, J. Dorado, A. Pazos, Curr. Comput. Aided Drug Des., 9 (2013) 108-117.
[15] J.S. Hamid, P. Hu, N.M. Roslin, V. Ling, C.M. Greenwood, J. Beyene, Human Genomics and Proteomics, 1 (2009).
[16] R. Jansen, N. Lan, J. Qian, M. Gerstein, Journal of structural and functional genomics, 2 (2002) 71-81.
[17] M. Weiner, T.E. Stump, C.M. Callahan, J.N. Lewis, C.J. McDonald, Int. J. Med. Inform., 71 (2003) 57-69.
[18] V.J. Zhu, M.J. Overhage, J. Egg, S.M. Downs, S.J. Grannis, J. Am. Med. Inform. Assoc., 16 (2009) 738-745.
[19] S. Gomatam, R. Carter, M. Ariet, G. Mitchell, Stat. Med., 21 (2002) 1485-1496.
[20] O.U. Press, 2013, pp. Definition of merge.
[21] Y.S.o. Medicine, 2013.
[22] J. Ma, N. Akhtar-Danesh, L. Dolovich, L. Thabane, BMC Med. Res. Methodol., 11 (2011) 18.
[23] M. Greiver, J. Barnsley, R. Glazier, B. Harvey, R. Moineddin, BMC Health Services Research, 12 (2012) 116.
[24] J. Braa, A. Heywood, S. Sahay, Bull. World Health Organ., 90 (2012) 379-384.
[25] L. Gu, R. Baxter, D. Vickers, C. Rainsford, CSIRO Mathematical and Information Sciences Technical Report, 3 (2003) 83.
[26] C.I. Neutel, Pharmacoepidemiol. Drug Saf., 6 (1998) 367-369.
[27] J.J. Berman, Artif. Intell. Med., 26 (2002) 25-36.
[28] D. Elgesem, Philosophical perspectives on computer-mediated communication, (1996) 45-66.
[29] G.B. Bell, A. Sethi, Communications of the ACM, 44 (2001) 83-88.
[30] G.W.M. Krysztof J. Cios, Artifical Intelligence in Medicine, 26 (2002) 1-24.
[31] M.B. Van Der Weyden, Med. J. Aust., 184 (2006) 430.
[32] N.H.a.M.R.C. Australian Government, NHMRC2007.
[33] Y. Reingewertz, Available at SSRN 2200023, (2013).
[34] R. Kammann, Hum. Factors, (1975).
[35] P. Dargan, C. Wallace, A. Jones, Emerg. Med. J., 19 (2002) 206-209.
[36] T. Crews, Computer Science Teaching Centre Digital Library, Western Kentucky University, USA. http://www.citidel.org/bitstream/10117/119/2/Visual. pdf, (2001).
[37] S. Institute, SAS 9. 3 Output Delivery System: User's Guide, Sas Institute2011.
[38] L. StataCorp, Stata Data Management: reference manual: release 12, StataCorp LP2011.

[39] N.E.M. Association, Digital Imaging and Communications in Medicine (DICOM), The Association1993.
[40] W.D. Bidgood, S.C. Horii, F.W. Prior, D.E. Van Syckle, J. Am. Med. Inform. Assoc., 4 (1997) 199-212.
[41] O.R. Zaiane, Principles of Knowledge Discovery in Databases, University of Alberta, 2013.
[42] O.R. Zaïane, CMPUT690, Dept. of Computing Science, University of Alberta, (1999).
[43] M.-S. Chen, J. Han, P.S. Yu, Knowledge and data Engineering, IEEE Transactions on, 8 (1996) 866-883.
[44] O.S. University, Ordination Methods for Ecologists, 2013.
[45] G. Tyburski, Trends L. Libr. Mgmt. & Tech., 6 (1994) 4.
[46] M.A. Hernández, S.J. Stolfo, ACM SIGMOD Record, ACM1995, pp. 127-138.
[47] M.A. Hernández, S.J. Stolfo, Data mining and knowledge discovery, 2 (1998) 9-37.
[48] M.A. Jaro, Stat. Med., 14 (2007) 491-498.
[49] Merging SAS Data Sets,SAS, 2013.
[50] Combining SAS Data Sets: Basic Concepts, SAS, 2013.
[51] H.B. Newcombe, J.M. Kennedy, Communications of the ACM, 5 (1962) 563-566.
[52] S. Gomatam, R. Carter, Technical Report, Department of Statistics, University of Florida1999.
[53] I.P. Fellegi, A.B. Sunter, Journal of the American Statistical Association, 64 (1969) 1183-1210.
[54] D. Dey, S. Sarkar, P. De, Knowledge and Data Engineering, IEEE Transactions on, 14 (2002) 567-582.
[55] M.U. Faculty of Human Sciences, 2013.
[56] W. Jenine Milum.
[57] M.J. Foley, Proceedings of the Twenty-Second Annual SAS Users Group International Conference1997, pp. 199-206.
[58] M. Scerbo, C. Dickstein, A.C. Wilson, Health Care Data and the SAS System, Sas Inst2001.
[59] SAS, in: S. Helper (Ed.)2013.
[60] P. Christen, K. Goiser, Proc. 4th AusDM 2005, (2005).
[61] F. Xu, M.-P. Austin, N. Reilly, L. Hilder, E.A. Sullivan, Arch Womens Ment Health, 15 (2012) 333-341.
[62] F. Xu, L. Hilder, M.-P. Austin, E.A. Sullivan, BMC Med. Res. Methodol., 12 (2012) 71.
[63] B.G. Tabachnick, L. Fidell, Using Multivariate Statistics: International Edition, Pearson2012.
[64] J.A. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, J.R. Carpenter, BMJ: British Medical Journal, 338 (2009).
[65] A.M. Wood, I.R. White, S.G. Thompson, Clinical trials, 1 (2004) 368-376.
[66] K.I. Penny, I. Atkinson, J. Clin. Nurs., 21 (2012) 2722-2729.
[67] A.C. Acock, Journal of Marriage and Family, 67 (2005) 1012-1028.
[68] S.R. Wisniewski, A.C. Leon, M.W. Otto, M.H. Trivedi, Biol Psychiatry, 59 (2006) 997-1000.
[69] J.L. Schafer, J.W. Graham, Psychol. Methods, 7 (2002) 147-177.
[70] X. Yan, S. Lee, N. Li, J. Biopharm. Stat., 19 (2009) 1085-1098.
[71] P.L. Roth, Pers. Psychol., 47 (1994) 537-560.
[72] I. Eekhout, R.M. de Boer, J.W. Twisk, H.C. de Vet, M.W. Heymans, Epidemiology, 23 (2012) 729-732.
[73] A. Pickles, Encyclopedia of social measurement, (2005) 689-694.
[74] B. Roure, D. Baurain, H. Philippe, Mol. Biol. Evol., 30 (2013) 197-214.
[75] F.M. Shrive, H. Stuart, H. Quan, W.A. Ghali, BMC Med. Res. Methodol., 6 (2006) 57.
[76] T.W. Anderson, Journal of the american Statistical Association, 52 (1957) 200-203.
[77] H. Hartley, R. Hocking, Biometrics, (1971) 783-823.
[78] D.B. Rubin, Biometrika, 63 (1976) 581-592.
[79] D.B. Rubin, Journal of the Royal Statistical Society. Series B (Methodological), (1976) 270-274.
[80] C.f.D.C.a.P.N.C.f.H. Statistics, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES 2013.

Figures

Table 1 Overview of the DIPIT ten-step protocol for data integration

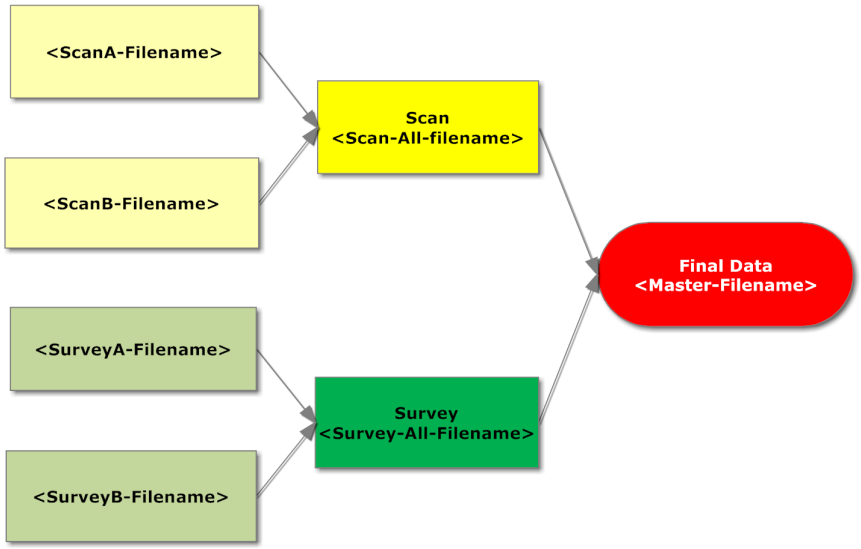
DIPIT STEP	Action	Strategy	Standard
1	Define the data requirements	• Define research hypotheses • Establish files to integrate • Assess data quality	Documentation of research hypotheses, files needed to integrate and data quality issues
2	Establish ethical, legal and privacy issues	Establish ethical, legal and privacy issues for each data file to integrate	Documentation of standards met
3	Order the files to integrate	Set up a flowchart for all files to be integrated, incorporating all file names	Flowchart of file hierarchy
4	Establish the file formats	Amend the flowchart in step 3 to document the file format for each file integrated and the final master file	Inclusion of all file formats in flowchart
5	Define the variables of interest	Create a table containing the variable of interest for research containing as a minimum: • Final variable name • Original variable name • Source file of variable • Preliminary file(s) for variable • Description of variable	Table of variables of interest for research incorporating a standard naming format, structured order and identification of file source.
6	Set up link(s) for integration	Create a table containing the variable(s) links and linkage method(s) used containing as a minimum: • Link variable(s) • Method of linkage • Automation used (if applicable)	Table of data file links, variables used and linkage method
7	Document the integration path	Document the structure of the path take for integration to include as a minimum: • The integration of the primary files • The saving of the *Master* file format in a standard file naming structure • The variables of interest to be retained • The variables standard naming format • The merging of all files into the *Master* file • A log of statistics of the key variables, and missing data analysis	Documentation of path of data file integration hierarchy incorporating primary and secondary files, logs and naming convention
8	Flowchart the type of integration	Document on flowchart type of integration: • one-to-one • many-to-one • one-to many • many-to-many	Method of integration included in flowchart and linkages used
9	Document the integration outcome	Define linkage quality measure. Table of mismatches of records by variable to contain as a minimum: • Variable name • Source of mismatch • Reason for mismatch	Documentation of degree of variable mismatches (e.g. log): which variables, percentage matched/mismatched. Document linkage quality measure (e.g. F-measure graphs).
10	Check variables and missing data	Initial data inspection to include as a minimum: • Analysis of key variable(s) • Missing data analysis	Document initial investigation of variables. Define minimum percentage of missing data acceptable for research based on industry convention and document future handling of missing data

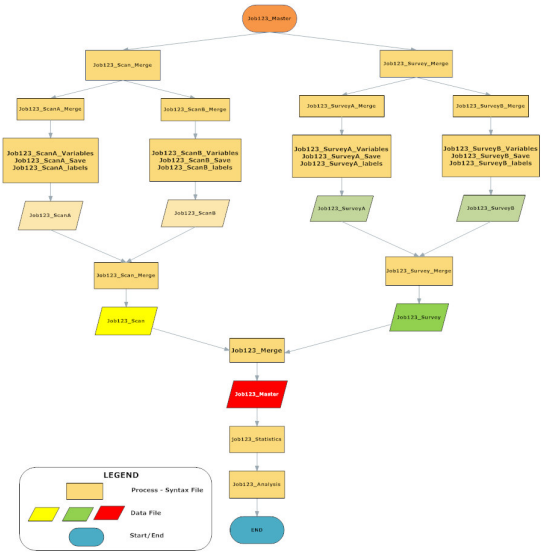
Table 2 Example Table of variable translations. File extensions included where applicable (e.g. ".dta" for Stata file format

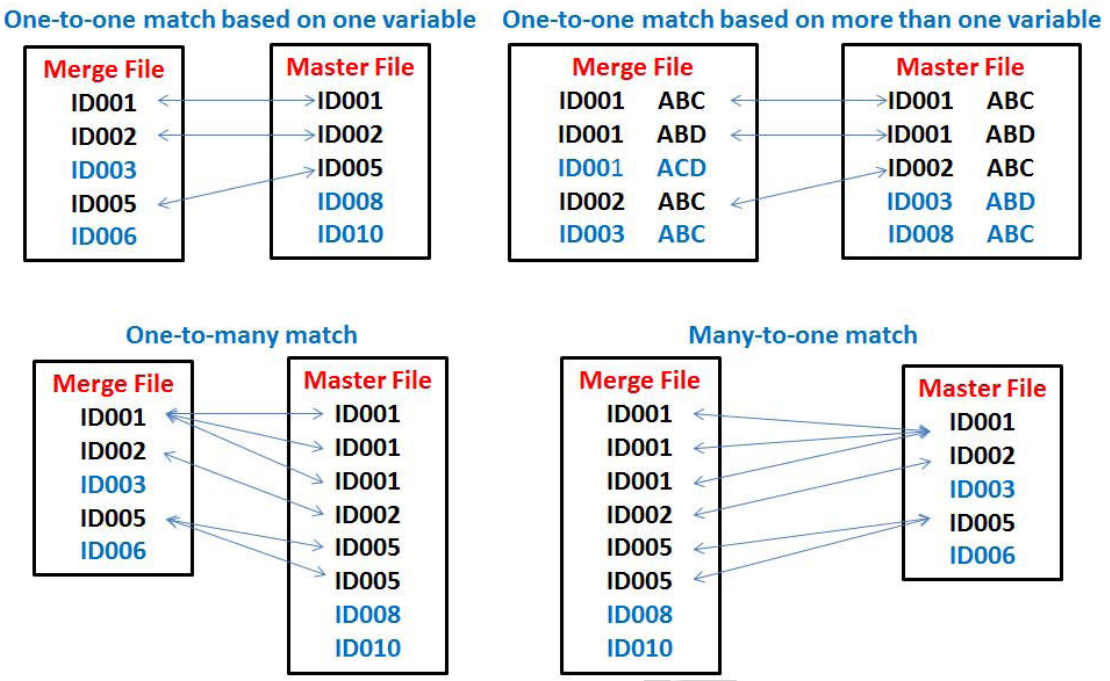
Final name(s)	Original name(s)	Source File(s)	Preliminary File(s)	Description
ID	ID	ALL	ALL	Unique patient identifier
Scan_larm	Larm	ScanA.csv	ScanA.dta	Scan of left arm
Scan_Rarm	Rarm	ScanA.csv	ScanA.dta	Scan of right arm
Scan_mri	Mri	ScanB.txt	ScanB.dta	MRI
Suvery_sex	Gender	SurveyA.csv	SurveyA.dta	Gender of patient
Survey_dob	DOB	SurveyA.csv	SurveyA.dta	Date of birth
Survey_Q10a to Survey_Q10s	Q10a to Q10s	SurveyB.sav	SurveyB.dta	Q10 Physical exercise items

Table 3 DIPIT Table using NHANES data integration

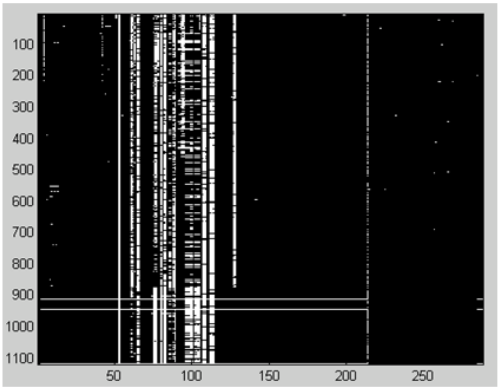
DIPIT STEP	Action	Strategy	Standard
1	Define the data requirements	• Define research hypotheses • Establish files to integrate • Assess data quality	To establish most important demographic & laboratory predictors for depression. All NHANES data files for the years 2005-2008 with consistent variables to be integrated. 80 data files were integrated. Data quality high as pre-processed prior to public access.
2	Establish ethical, legal and privacy issues	Establish ethical, legal and privacy issues for each data file to integrate	NHANES 2005-2006: Protocol #2005-06 NHANES 2007-2008: Continuation of Protocol #2005-06
3	Order the files to integrate	Set up a flowchart for all files to be integrated, incorporating all file names	Flowchart *NHANES-JobDep1-FL001.ppt* in Microsoft PowerPoint format
4	Establish the file formats	Amend the flowchart in step 3 to document the file format for each file integrated and the final master file	All original files were SAS files, initially transferred into Stata format for integration. Consistent primary file names between SAS and Stata.
5	Define the variables of interest	Create a table containing the variable of interest for research containing as a minimum: • Final variable name • Original variable name • Source file of variable • Preliminary file(s) for variable • Description of variable	Microsoft Excel file named *NHANES-JobDep1-EXC001.xlsx* of variables excluded
6	Set up link(s) for integration	Create a table containing the variable(s) links and linkage method(s) used containing as a minimum: • Link variable(s) • Method of linkage • Automation used (if applicable)	Microsoft Excel file named *NHANES-JobDep1-LNK001.xlsx* of data file links and linkage method.
7	Document the integration path	Document the structure of the path take for integration to include as a minimum: • The integration of the primary files • The saving of the *Master* file format in a standard file naming structure • The variables of interest to be retained • The variables standard naming format • The merging of all files into the *Master* file • A log of statistics of the key variables, and missing data analysis	Documentation of path of data file Dintegration stored in Stata syntax *.do* files with one "Master" integration *.do* file named *NHANES-JobDep1-Master-Integrate.do* used to run all subsequent syntax files. A log file of all key variable and missing data stored in the name *NHANES-JobDep1-Stats1.txt.*
8	Flowchart the type of integration	Document on flowchart type of integration: • one-to-one • many-to-one • one-to many • many-to-many	All integration one-to-one merge based on unique sequential number for variable *seqn.*
9	Document the integration outcome	Define linkage quality measure. Table of mismatches of records by variable to contain as a minimum: • Variable name • Source of mismatch • Reason for mismatch	All merge variables contained details of mismatches and source at each stage of integration. Merge variables denoted by *merge_[fileID]* in Master Stata data file.
10	Check variables and missing data	Initial data inspection to include as a minimum: • Analysis of key variable(s) • Missing data analysis	Only those with a depression score were included and criterion >50% non-missing. Microsoft Excel file named *NHANES-JobDep1-VAR00*1.xlsx for final set of integrated variables.
Note: Actual file names altered from original.			







_pattern	_mv	_freq
++++++++++_+++_..._+++++++++++++++++_..._++++++++++++++++++++	6	309
++++++_+++_+++_..._+++++++++++++++++_..._++++++++++++++++++++	7	243
++++++++++++++++++++++++++++++++++++_..._++++++++++++++++++++	2	84
++	0	73
++++++_++++++++++++++++++++++++++++_..._++++++++++++++++++++	3	50
+++_++++++_+++_..._+++++++++++++++++_..._++++++++++++++++++++	7	46
+++++++++_++_..._+++++++++++++++++_..._++++++++++++++++++++	9	40
+++_++_+++_+++_..._+++++++++++++++++_..._++++++++++++++++++++	8	26
++++++_++++++_..._+++++++++++++++++_..._++++++++++++++++++++	6	22
----------	58	22
++++++++++_+++_..._++	4	20
++++++_+++_+++_..._++	5	19
+++++++++_++_..._++	6	19
++++++++++++++++++++++++++++++++++++_..._++++++++++++++++++++	3	16
++++++++++++++++_..._+++++++++++++++++_..._++++++++++++++++++++	5	14
++++++++++++++_..._++	3	11
+++_++	1	7
++++++++++++++_..._++	4	7
++++++++++++++_..._++++++++++++++++_..._++++++++++++++++++++	7	6
+++_++++++++++++++++++++++++++++++++_..._++++++++++++++++++++	3	5
+++_++_++++++++++++++++++++++++++++++_..._++++++++++++++++++++	4	4
+++++++++_++_..._++++++++++++++++++++_..._++++++++++++++++++++	8	4
++++++++++_++_..._++++++++++++++++++++_..._++++++++++++++++++++	7	3
+++_++_+++++++_..._++++++++++++++++++++_..._++++++++++++++++++++	7	3
++++++++++_++_..._++++++++++++++++_..._++++++++++++++++++++	8	3
++++++++++_++_..._++	5	2
+++_++++++_+++_..._++	5	2
+++++++++_+++_..._++++++++++++++++_..._++++++++++++++++++++	8	2
++++++++++_++	1	1
++++_++	1	1
++++++++++_..._++	2	1
++++++++++_++_..._++	3	1
+++++++++_++_..._++	4	1
+++_++++++++++_..._++	4	1
++++++++++_+++_..._++	5	1
++++++++++_++_..._++++++++++++++++_..._++++++++++++++++++++	7	1
+++++++++_++_..._++++++++++++++++_..._++++++++++++++++++++	8	1
+++_+++++_+++_..._++++++++++++++++_..._++++++++++++++++++++	8	1
+++_+++++_++_..._++++++++++++++++_..._++++++++++++++++++++	10	1



Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers.

[MAX 5 POINTS, 85 CHARACTERS PER POINT]

- Currently no documented protocols exist for best practice integration of data files
- Poor quality integration processes cause errors and loss of confidence in the data
- DIPIT is a systematic approach for integrating multiple heterogeneous data files
- DIPIT is designed to minimise errors and streamline the integration process