PREDICT RA Workshop

Trial Data Management

Luke Stevens
Data Management Coordinator
Clinical Epidemiology and Biostatistics Unit
Murdoch Childrens Research Institute
www.mcri.edu.au
luke.stevens@mcri.edu.au

Topics

- Primary Principles
- Data Collection
- Databases
- Development and testing
- Managing live data
- Using your data

Primary Principles

- Good quality data prerequisite for good quality results
- Garbage in – garbage out
- Good documentation
- Good organisation
- Good procedures
- *** REPRODUCIBLE RESULTS ***

Reproducibility Throughout

- Study manual, protocol
    - Documenting how study will be run
    - Data collection procedures
- Standard Operating Procedures
    - Specific instructions for how tasks are to be completed
- Data collection, data entry, data custodianship
    - Audit trail (paper or electronic)
    - Security and permissions controls
- Analysis
    - File management: folders, naming, version control
    - Cleaning and analysis scripts

Data Collection

- Paper or electronic?
    - Technological considerations – what is available?
    - Practical considerations – what will work best in the collection setting?
- Paper
    - Easy and convenient in face-to-face setting
    - Hard-copy source document
    - Requires handling and storage
- Electronic
    - No additional data entry time
    - Validation at source – much less cleaning time
    - Higher training requirements

Trial Databases

- Choosing what to use
    - Audit trail
    - User permission controls
    - Secure storage
    - Data quality measures
    - Data export to statistical software
    - Report capabilities
    - Functionality (e.g. web access, data queries, monitoring, randomisation)
- Options?
    - Not recommended: Excel, stats packages, Access
    - Better options: EpiData, REDCap, WebSpirit

Databases: Can I Use Excel?

- Microsoft Excel is a spreadsheet, not a database
- Use only if you do not care about your data’s:
    - Integrity
        - Move data across records and columns
        - No audit trail
    - Quality
        - No data type enforcement
        - No range checks or cross-field validation
    - Security
        - No user or permission management
        - File-based, so manual version control and backup
    - Usability
        - No metadata means no direct export to stats package

Databases: Why not just use my stats package?

- Preserve record integrity, columns have fixed data type
- Do not offer other essential or desirable features
    - Audit trial
    - User access and permission controls
    - Validation checks upon entry (except data type)
    - File-based, so manual version control and backup
    - Workflow functionality (queries, randomisation etc.)
- Use the correct tool for the job. There are better options.

Databases: Microsoft Access / FileMaker Pro

- Both combine relational database (back-end) with a user interface (front-end):
    - File-based
    - Vast scope for bespoke forms and functionality
    - Specific programming expertise required
- Recommend use:
    - Only when your team has the necessary skills
    - As a supplementary system e.g. for letters, reports
    - *Not* recommended as primary database for your research data

Databases: EpiData Entry

- Features:
  - Free software: download from epidata.dk, install
  - File-based datasets
  - Excellent data validation
  - Exports datasets directly to stats package binary files
- Recommend use:
  - Single user
  - Simple, single form
  - Data entry of data collected on paper

Databases: WebSpirit

- Features:
    - Web-based – accessible from anywhere
    - Trial workflows (monitoring, record sign-off etc.)
    - Scope for complex forms and event schedules
    - Audit trail, user permissions, data validation
- Recommend use:
    - Robust security and data quality
    - Trials
    - Institution member of PTNA

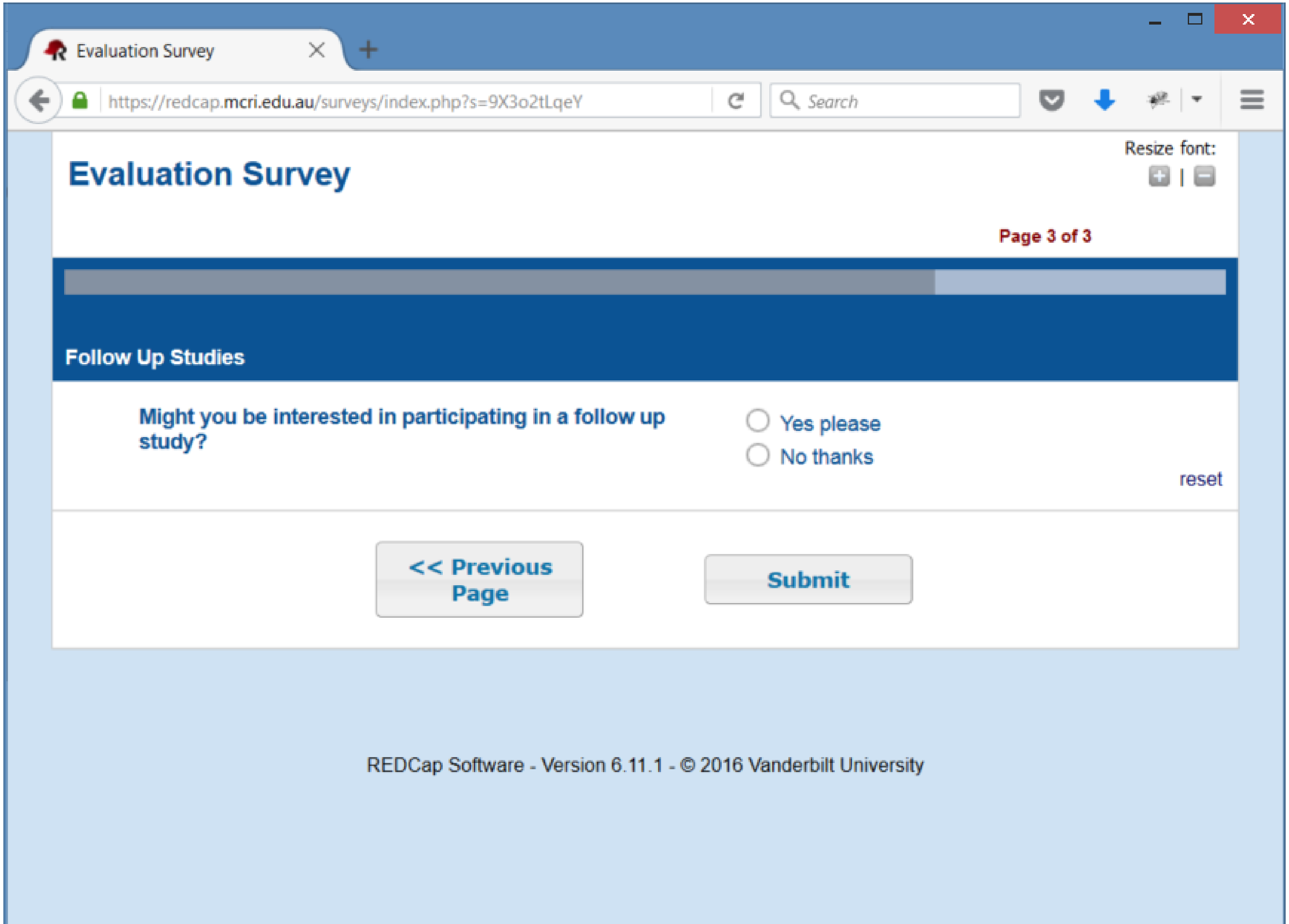
Databases: REDCap

- Features:
    - Web-based – accessible from anywhere
    - Authenticated data entry and online survey forms
    - Flexible, broad range of functionality
    - Audit trail, user permissions, data validation
- Recommend use:
    - Robust security and data quality
    - Rapid development, piloting
    - Scope for customisation
    - Application hosted by your institution

Database Development: Survey Form Design

- Design forms mindful of data entry method
- Participant-completed survey forms will be viewed by each participant only once
    - Use radio buttons rather than drop-down lists so that all options are visible without having to select the field
    - Break the form into small sections, with section-per-page, to capture partial responses
    - You can be liberal with explanatory text and design elements (e.g. images)
    - Provide “do not wish to answer” options (where appropriate)

Database Development: Survey Form Design



Database Development: Data Entry Form Design

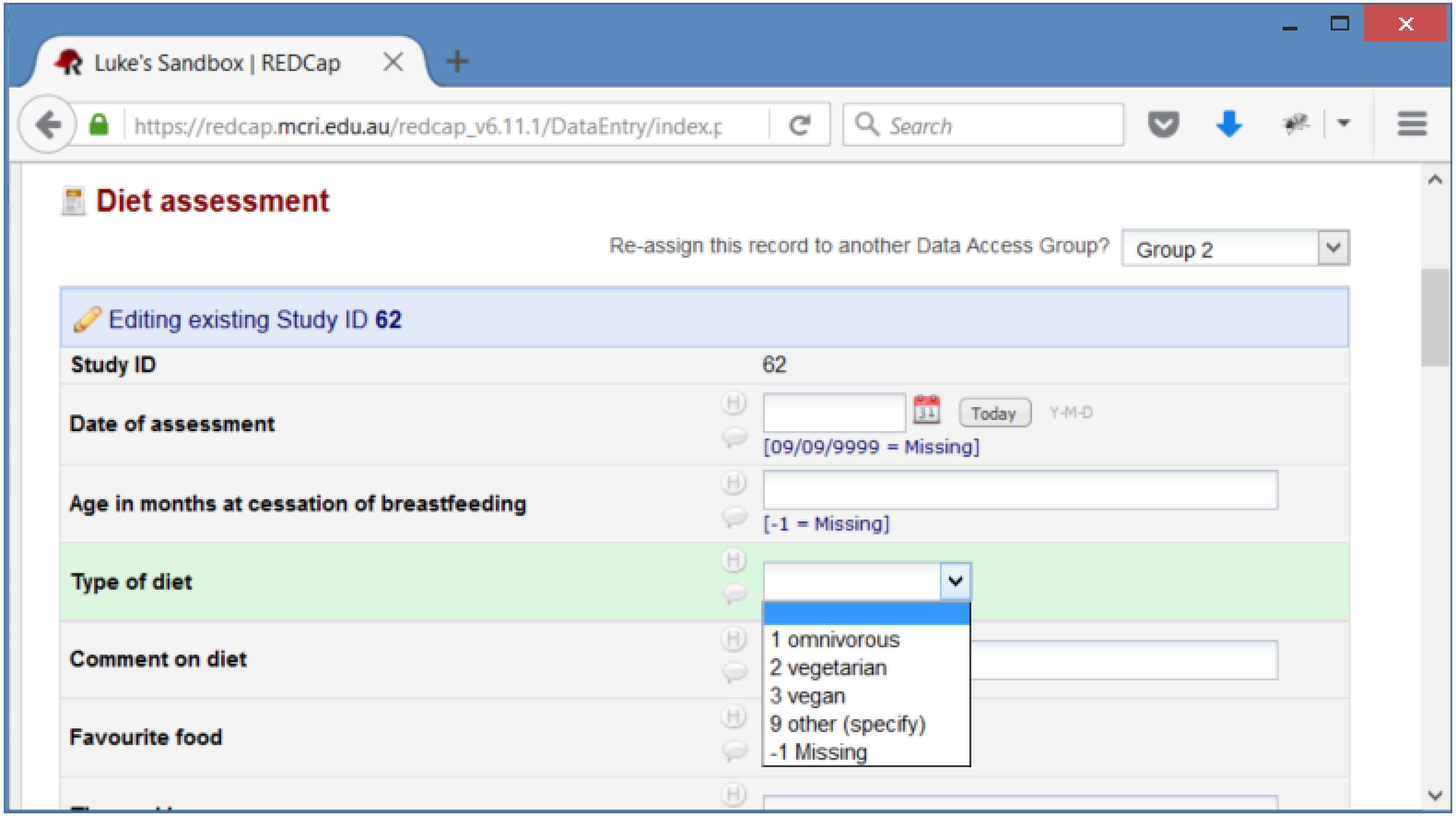
- Forms for entry by data entry person viewed many times
  - Use drop-down lists rather than radio buttons so that options can be selected with fewest key-strokes
  - Begin value labels with the corresponding value. E.g.:

    0, 0 No

    1, 1 Yes

    Data can be entered using number then tab to next.
  - Make all fields mandatory. Include codes for missing values for every field – nothing is mandatory on paper!
  - Lay out the data entry form matching the sequence of how the paper form will be read as closely as possible
  - Design should be as simple and uncluttered as possible (go easy on images, text styles etc.)

Database Development: Data Entry Form Design



Database Development: Testing

- Test thoroughly
    - Ensure data entry forms function as required
    - Ensure other project configurations (e.g. user permissions, automated emails, randomisation) are set up correctly and appropriately
- User training
    - Users must become familiar with the navigating database
    - Each person using the database must know how to perform their tasks correctly
- Piloting
    - Piloting your forms with people like your participants – not just the study team – is invaluable
    - Even if on paper, a good test of data entry forms

Database Development: Access Controls

- "Principle of least privilege"
  - Users need access to just those functions and data they require to perform their tasks – no more
  - Simplify training
  - Reduce scope for error
- User permissions/access considerations
  - Define user types/roles according to tasks
  - Study-level vs. site-level users
- Participant identifiers
  - Ensure participant identifiers are not accessible to any user that does not need to see them
  - Participant tracking data and study data may be separated into different databases
  - Be careful with free-text fields

Managing a Live Database: Data Changes

- Inevitable!
- Restrict users able to perform
- Audit trail
    - Essential!
    - Include reason for change
    - On paper if necessary
- Have a SOP
    - How are required changes to be identified?
    - How is the appropriate resolution determined?
    - Who will carry out each step
    - Consider a “Data Issues Log” to document:
        - Description of the problem
        - Suggested resolution
        - Approval and implementation of the resolution

Managing a Live Database: Design Changes

- Also inevitable!
- Adding new data collection elements less problematic than removing or altering
- If removing, consider "retiring" fields by hiding them rather than deleting
- If altering, do not alter the meaning of a variable or value
  - For example, changing the label for option 3:

1, Thing 1	1, Thing 1
2, Thing 2	2, Thing 2
3, Other	3, Thing 3
	4, Other

Any records where option 3, "Other", was selected will now be labelled "Thing 3"

Using Your Data: Export from Database

- Make sure you know how data is exported
  - EpiData exports stats package binary files (.dta, .sav)
  - REDCap, WebSpirit give you raw data in .csv format plus a script (.do, .sps) that reads in the data and labels variables and values to generate a dataset
- Ensure you know what the data will look like, e.g.
  - REDCap's row-per-participant-per-event longitudinal data
  - WebSpirit's nr and enr columns for repeated forms or groups/tables
- Data access process and requirements
  - Who has or can gain access to the data
  - What data is accessible – protect identifiers
  - "Statistician" or "Data Manager" roles
  - A "Data Access Request" SOP

Using Your Data: Preparation for Analyses

- File management
  - Directory structure conventions
    - Indicate purpose, version
  - File naming conventions
    - Indicate purpose, version, date

- All cleaning and dataset preparation should be performed using stats package scripts
  - Stata .do files, SPSS .sav files
  - No ad hoc point and click from menus
  - Preserve raw source data
  - Document reasons for each operation

```
/* Ensure withdrawn participants are dropped */
drop if record_id = ‘1424’ | record_id = ‘5460’
```

  - Remember reproducibility!

Summary

- Plan, document and organise your data processes
- Choose an appropriate database for each data collection
- Consider collection mechanisms and design forms accordingly
- Thorough testing and user training
- Implement data quality control mechanisms
- Implement data access controls and procedures
- Be systematic with your data processing and analyses
- Reproducibility is the goal!