

---

# Data Management in Clinical Trials

---

Kit Howard, MS  
Kestrel Consultants, Inc.  
17 March 2005

---

# Premise

- Defining the right clinical question and the appropriate statistical approach are critical to the success of a clinical study.
  - While often overlooked, it is just as critical to understand what data to collect, how to collect them, how to ensure they are of high quality, and how to 'database' them.
  - This talk will present an overview of these topics, along with some practical tips for ensuring that the study collects the right data in the right way to answer the study question.
-

---

# Outline

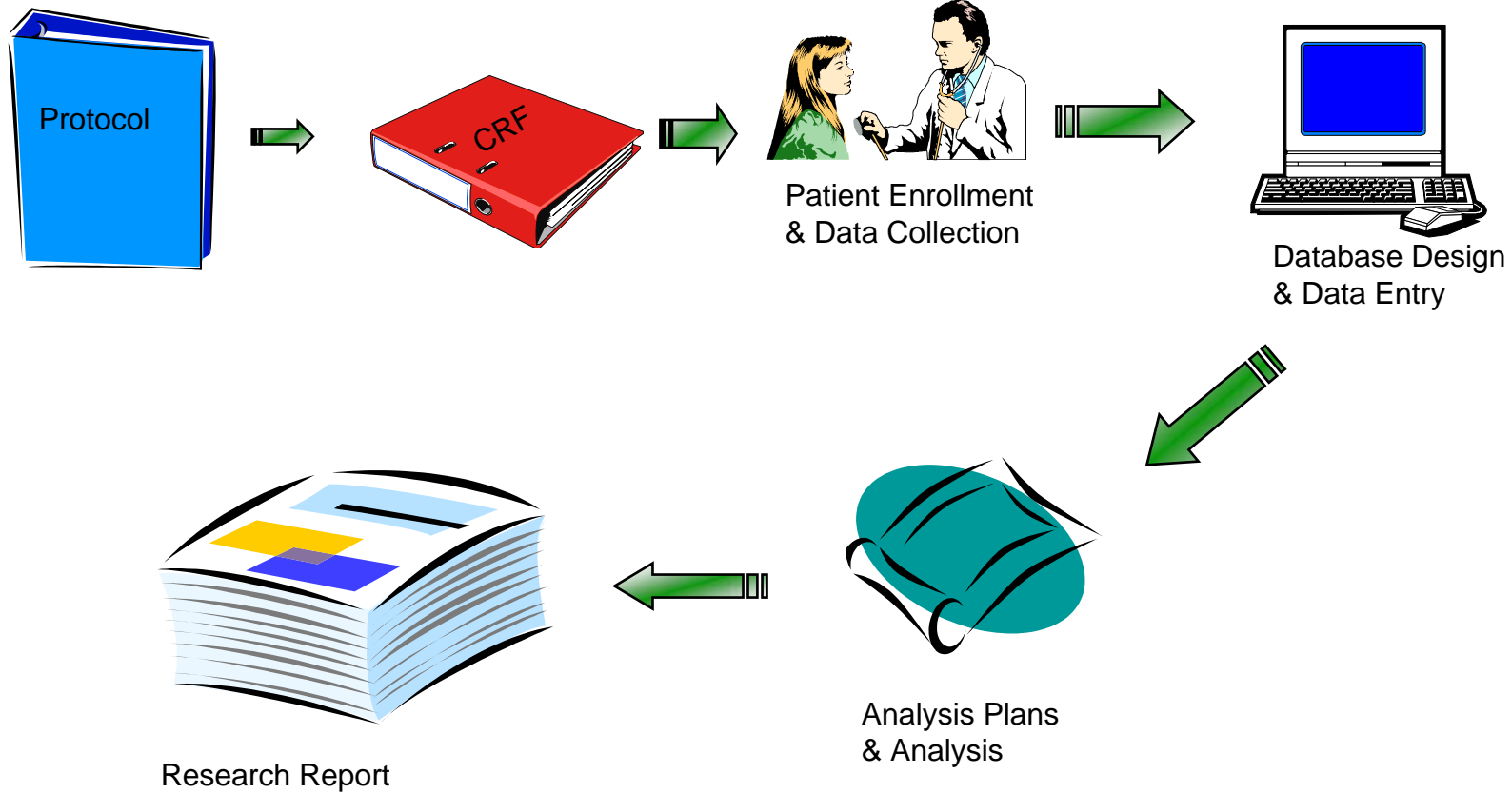
- Protocol definition
  - Question definition
  - Terms definition
  - Analysis approach
  - Summary & analysis layouts
  - Data collection
  - Data coding
  - Data quality
  - Database design and development
  - Concluding thoughts
-

---

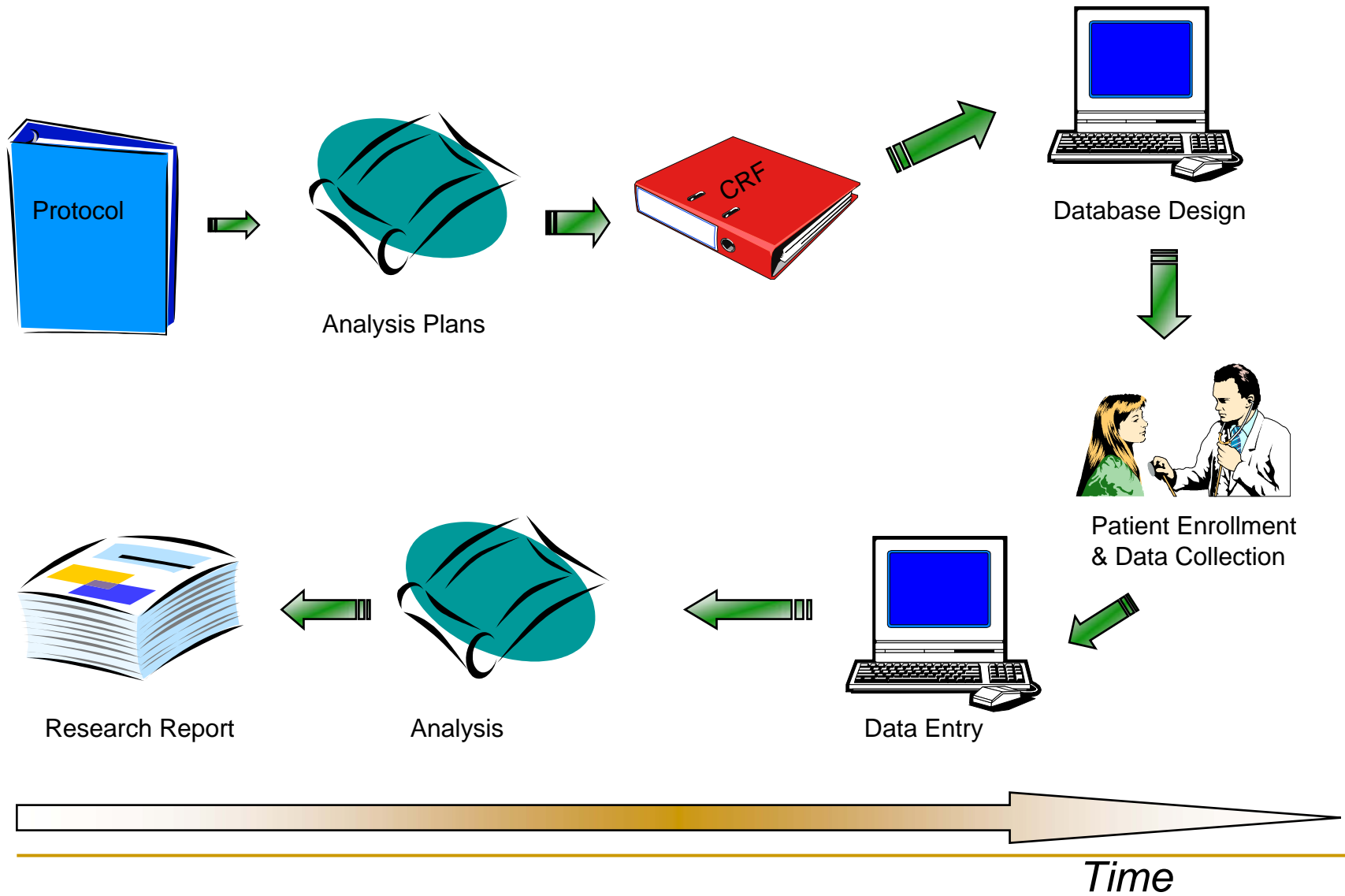
# Context for this presentation

- Who is doing the study?
  - Driver: ensure that all the data are collected in the right way the first time
  - Not a programming class
-

# Usual Practice: Chronological



# Alternative Practice



# Efficacy & Safety of Compound X

- Compound X is a potential treatment for facial blackheads on the nose, completely and painlessly eliminating them in just hours with just a single application
- Double blind, randomized trial comparing Compound X and an inert cream, and the randomization determines whether Compound X is applied to the right or left side of the nose
- Compare left and right side (internal control)
- Apply cream during office visit
- Return 24 hours later for assessment
- Outcome is success, partial success or failure
- Trial considered a success if achieve at least partial success in at least 60% of the patients, with minor or no side effects
- Power calculations indicate that at least 50 subjects are needed

---

# Operationalizing the Protocol

- Have written the grant proposal, received the grant
  - Now need refine and operationalize the protocol
    - What are the precise questions want to answer
    - How define the terms
    - How need to present what data to test hypothesis
    - Where and how will get the data
    - How do the data need to be structured in order to use them as planned
    - How to collect the data
    - How to assess its consistency and quality
    - How to “database” the data
-

---

# Defining the Questions Precisely

- What is the demographic profile of the study subjects
  - Does the treatment work, based on hypothesis – how will this be assessed
  - Do any of the efficacy or safety parameters vary by age, sex, weight, race, or skin type
  - What other factors beyond demographic variables might affect the results
  - What side effects are observed, how frequently and where do they occur, and how significant are they
-

---

# Defining the terms

- Define 'success', 'partial success', 'failure' (# blackheads per given area?)
  - Decide how/when to determine baseline, how to record that
  - Define skin type, race
-

---

# Presenting & Analyzing the Data

- For each question, or set of questions, design a table layout that lets you assess the answer. This will also help to define the data you need to collect
    - Analyses (list equation(s), tests and/or variables will need)
    - Summary table(s) (do sample layouts)
    - Listings (do sample layouts)
-

---

# Data Categories

- “Header” information
  - Efficacy parameters
  - Safety parameters
  - Demography
  - Physical exam
  - Treatment information
-

---

# “Header” Information

- Study identifier on every page
  - Subject identifier on every page
  - Dates
    - Visit date
    - Assessment or event date
-

---

# Efficacy

- How to assess the treatment
    - Full success: complete elimination of blackheads
    - Partial: elimination of at least 50% of blackheads
    - Failure: elimination of less than 50% of blackheads
  - # blackheads per defined surface area (e.g., 1 square inch to right and left of nose center)
    - Mark the areas? How know that will assess same area?
    - Does the average size of the blackhead matter?
      - Measurement?
      - Small/med/large? Definition?
-

# Sample Data Listing

Study Title  
Subject Listing of Number of Blackheads

| Treatment Group  | Subject Number | Study Day | # Blackheads Left | # Blackheads Right | Difference (%) |
|------------------|----------------|-----------|-------------------|--------------------|----------------|
| Compound X Right | 1              | xx        |                   |                    |                |
|                  |                | xx        |                   |                    |                |
| Compound X Left  | 3              | xx        |                   |                    |                |
|                  |                | xx        |                   |                    |                |

# Sample Summary Table

| Treatment Result | Study Title                     |           |                    |           |
|------------------|---------------------------------|-----------|--------------------|-----------|
|                  | Summary of Treatment Assessment |           |                    |           |
|                  | Subject Population              |           | Subject Population |           |
|                  | Placebo                         |           | Compound X         |           |
|                  | N                               | Mean (SD) | N                  | Mean (SD) |
| Full Success     |                                 |           |                    |           |
| Partial Success  |                                 |           |                    |           |
| Failure          |                                 |           |                    |           |
| Total            |                                 |           |                    |           |

---

# Safety

- Adverse events, & associated parameters
    - Name, duration, severity
  - Defining duration (start/stop date, time)
  - Defining severity (mild, mod, severe – meaning?)
  - Specific collection of skin reactions?
-

---

# Demography

- Sex – genetic? Physiological?
  - Race – definition?
  - Age (birth date & study date?)
  - Hormonal status
-

---

# Physical exam

- Blood pressure
  - Heart rate (units?)
  - General physical (body systems?)
  - Height/weight (units?)
  - Skin type (definition?)
-

---

# Treatment

- Record of which treatment arm received
- Time of dose?

---

# Good Data Capture Principles

---

---

# Good Data Capture Principles (1)

- Never collect the same data in more than one place
    - This invites error in one place or the other
    - Ensures the need for additional data cleaning
  - Do not collect leading or open-ended questions if:
    - Data are needed for summarization
    - Need detailed information & consistency
  - Collect data in a fashion that allows for the most efficient computerization
    - Use pre-codes for data that may need to be summarized or searched on
    - Use checkboxes whenever possible
    - Use a straight forward layout
-

---

# Good Data Capture Principles (2)

- Provide units to help ensure comparable values. If necessary, provide a choice of units.
  - As a general rule, collect raw data and only do calculations manually on data needed during the study.
  - Choices provided for each question should include all options that may be encountered and/or have 'other' category for unexpected.
  - “Absence of evidence is not evidence of absence”
    - If findings were not present or test was not done, there should be a way to indicate this.
    - Include 'None' or 'Not Done' as choices
    - Do not assume that because a data capture form is blank, there are no findings – check the source if possible
-

---

# Good Data Capture Principles (3)

- Be aware of when data points are required and what makes logical sense when grouping data modules or data points on the same page

Design data collection tools with anticipated data collection time points in mind

- Complete and accurate instructions for data capture should be provided in the notebook
-

---

# Good Data Capture Principles (4)

- Ensure enough space is provided to capture expected data
  - Provide a pattern of completion that is consistent with how the data are being generated
  - Ensure that forms are not 'too' busy and are readable for all users
-

---

# Coding Dictionaries

- In order to summarize data, it must be countable
  - In order to be countable, the values must be identical
  - Text data are impossible to summarize as they cannot be relied upon to be identical
  - All data to be summarized must be coded
  - Coding involves assigning standard numbers or text to common data
-

---

# Coding Process

- Coding happens at 2 time points
    - Predefined limited list of values: codes created when CRF is developed
    - Very large or infinite set of possible values: codes assigned by sponsor after data are entered
  - For large sets of values, use coding dictionaries
    - Dictionaries are structured computer files containing recorded terms, associated preferred text terms, and assigned code numbers
    - Most dictionaries categorize items by type, body system, use, or some other method of organization
    - CDM usually initially assigns the codes, with review by a clinician, nurse, or other medically trained individual
  - Data that are most commonly coded using dictionaries are AEs and medications
  - Other data may also be coded, depending upon the study analysis requirements
-

---

# Auto-encoding Systems

- Occasionally data may be coded by hand
  - Often data are coded using a program to compare the CRF text to the recorded terms in the dictionary
  - When a match is found, the system automatically assigns the preferred term and code number
  - All terms that have no match are output for investigation
    - May just require spelling changes, correcting abbreviations or acronyms, splitting multiple events listed as one
    - Some may require clarification
-

---

# Types of Coding Systems

- Many different dictionaries exist, each created for different needs
  - Dictionary chosen will depend upon the analysis needs
  - May have access to a commonly used dictionary
  - May be dictated by regulatory requirements
  - Some commonly used dictionaries
    - COSTART
    - WHO Drug
    - MedDRA
    - LOINC
    - WHOART
    - ICD-9-CM
    - SNOMED
-

---

# Clinical Database Development

- Clinical data are generally entered into electronic databases to facilitate analysis
  - Database systems can include
    - Excel or other spreadsheet
    - Relational databases (e.g., Access, Oracle)
    - Non-relational databases (e.g., SAS files)
  - What is used depends upon the volume of data expected and what will be done with them
-

---

# Database Development Steps

- Data output and analysis needs are defined
  - Data to be stored are identified
    - CRF items
    - Calculated fields
    - Key or identifier fields
  - Details for each data item are defined
    - Name, length, type
  - Annotated CRFs are created
  - Code lists are developed
-

---

# Database Development Steps

- Files are designed and built
  - Data entry screens that mimic the CRFs are designed and built and linked to the data files
  - Test data for every field are entered
  - Test data are exported and errors are corrected
  - Database is put into production
-

---

# Data Entry Guidelines

- A set of instructions is developed for how to enter data
  - Intended to ensure that all operators enter the data the same way
  - E.g., what to do if a field is illegible, or a required field is blank
-

---

# Data Entry

- Common methods
    - Single key entry
      - One person enters the data
      - Another person prints the results and compares with source documents
    - Double key entry
      - One person enters the data, then another person enters the same data into a mirror screen and the system compares the values real-time
      - Discrepancies are identified to the second operator as they happen, and they can be corrected immediately
    - Independent double key entry
      - Similar to double key entry except that the second entry is independently done and a program is run later to compare the two sets of entries
-

---

# Data Quality

---

---

# Definition

- Quality data are “data that support conclusions and interpretations equivalent to those derived from error free data”\*
- Challenge is to know what “error free data” are

\* *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making*, Institute of Medicine Roundtable on Research and Development of Drugs, Biologics, and Medical Devices Jonathan R. Davis, Vivian P. Nolan, Janet Woodcock, and Ronald W. Estabrook, *Editors*

---

---

# Definition

- Larry English additionally defines *inherent* and *pragmatic* quality
  - Inherent quality refers to the “correctness or accuracy of data”
  - Pragmatic quality is “the value that accurate data has in supporting the work of the enterprise”\*

---

# Definition

- Establishing inherent data quality has two elements
    - Accuracy of data entry, aka data verification
      - Do the data in the database accurately reflect the data on the CRFs
    - Correctness of data, aka data validation
      - Do the data reflect the reality of what happened, and do they make logical sense
-

---

# Data Management Plan

- Data Management Plan
    - The document that outlines all the data management practices for a given study
      - CRF completion guidelines
      - CRF tracking instructions
      - Data entry guidelines
      - The data querying process
      - How corrections will be made
    - Data quality is defined in the Data Management Plan
      - The tools and processes to be used to verify and validate the data
      - The level of quality required and its assessment
      - Who can make what kinds of corrections
  - The Plan changes between studies depending upon the data being collected and the uses to which it will be put
-

---

# “Clean Data” Rules

- “Clean Data” rules come from several sources
    - compatibility with life
    - statistical assumptions
    - protocol requirements
    - regulatory requirements
  - Data that violate the rules aren’t always changed
    - Sometimes it’s what really happened
-

---

# Data Validation Checks

- Purpose
    - To ensure that data are logical and meet statistical assumptions
    - To flag potential problems for further investigation
  - Types of Checks
    - Programmed/Automated Checks (aka edit checks)
      - Range Checks
      - Value Checks
      - Logical Checks
    - Manual Checks
      - Listing review
      - Statistical tools
    - Self Evident Changes (SECs)
-

---

# Database Completion Activities

- Data are entered and cleaned throughout the study
  - Dirty data may contain aberrant values that will distort a statistical analysis
  - To avoid this, prior to analysis data quality activities are conducted
    - Execute the data validation plan
    - Generate and resolve all queries
    - Ensure all coding has been performed
  - The quality control is performed and all issues are resolved
  - The database is 'released' for analysis, i.e., locked
  - To provide a stable dataset for analysis, the database is not changed after release except in extreme cases
-

---

# Breaking the Blind

- Once the database has been released, the statistician receives a file containing the treatment code
- This code reveals which patients were on what treatment
- The codes are loaded or entered into the database
- Now analysis programs can be run that group the patients by treatment to look for differences

Note that this is done after data cleaning is complete!

---

---

# Concluding thoughts

- Increasing rigor in documentation and ensuring quality and consistency should be exercised
    - the more people are involved in running the study
    - the longer the trial
    - the more data collected
    - the more subjects enrolled
  - Be wary of collecting info just because you can
    - Expensive
    - Increases chances of spurious results
    - Muddies the answer
-

---

Thank you.

---

Kit Howard

[Kit@KestrelConsultants.com](mailto:Kit@KestrelConsultants.com)

734-576-3031