Practical Elements of Database Design

NCIC CTG Course for New Investigators
August 9-12, 2011

Learning Objectives

At the end of the session the participant should be able to:

- identify principles of database design
- understand linkage between protocol and publication
- understand data element requirements
- be familiar with challenges and solutions in database design

Goals of CRF/Database Design

- Ensure that information from primary source records are recorded in a manner that allows for accurate processing, analysis, interpretation, reporting/ publication
  - Eligibility, treatment, AEs, endpoints
- Fulfills data collection requirements for regulatory and compliance purposes
- May also facilitate data exchange (meta-analysis)

Types of data collection

- *Prospective* clinical trial with pre-defined data collection
- *Retrospective* data abstraction for a defined research project

Case Report Forms/Database

- CRF is a *tool* to collect data
- May be electronic or paper based
- Standardized recording of information over time to produce robust, accurate, reproducible results
- Database is only as good as the quality of collected data!

Design Principle #1:
Protocol to Publication

- Consider content of the final publication at the time the protocol is being written
- Data to be collected in support of objectives should be described in the study protocol
- Data specified in protocol should then be collected on the CRF

Example: Published Baseline data in Melanoma RCT

Table 1. Baseline Characteristics of the Patients.*

Variable	Ipilimumab plus gp100 (N=403)	Ipilimumab Alone (N=137)	gp100 Alone (N=136)	Total (N=676)
Mean age — yr	55.6	56.8	57.4	56.2
Sex — no. (%)				
Male	247 (61.3)	81 (59.1)	73 (53.7)	401 (59.3)
Female	156 (38.7)	56 (40.9)	63 (46.3)	275 (40.7)
ECOG performance status — no. (%)†				
0	232 (57.6)	72 (52.6)	70 (51.5)	374 (55.3)
1	166 (41.2)	64 (46.7)	61 (44.9)	291 (43.0)
2	4 (1.0)	1 (0.7)	4 (2.9)	9 (1.3)
3	1 (0.2)	0	0	1 (0.1)
Unknown	0	0	1 (0.7)	1 (0.1)
M stage — no. (%)‡				
M0	5 (1.2)	1 (0.7)	4 (2.9)	10 (1.5)
M1a	37 (9.2)	14 (10.2)	11 (8.1)	62 (9.2)
M1b	76 (18.9)	22 (16.1)	23 (16.9)	121 (17.9)
M1c	285 (70.7)	100 (73.0)	98 (72.1)	483 (71.4)
Lactate dehydrogenase level — no. (%)				
≤Upper limit of the normal range	252 (62.5)	84 (61.3)	81 (59.6)	417 (61.7)
>Upper limit of the normal range	149 (37.0)	53 (38.7)	52 (38.2)	254 (37.6)
Unknown	2 (0.5)	0	3 (2.2)	5 (0.7)

Example: Published Toxicity Data in CRC RCT

Table 2. Adverse Events.

Event	Cetuximab plus Best Supportive Care (N = 288)				Best Supportive Care Alone (N = 274)				P Value
					number (percent)				
Grade 3 or higher with an incidence of ≥5%*									
Any adverse event	226 (78.5)				162 (59.1)				<0.001
Edema	15 (5.2)				16 (5.8)				0.85
Fatigue	95 (33.0)				71 (25.9)				0.09
Anorexia	24 (8.3)				16 (5.8)				0.32
Constipation	10 (3.5)				13 (4.7)				0.53
Nausea	16 (5.6)				15 (5.5)				1.00
Vomiting	16 (5.6)				15 (5.5)				1.00
Non-neutropenic infection	37 (12.8)				15 (5.5)				0.003
Confusion	16 (5.6)				6 (2.2)				0.05
Abdominal pain	38 (13.2)				43 (15.7)				0.40
Other pain†	43 (14.9)				20 (7.3)				0.005
Dyspnea	47 (16.3)				34 (12.4)				0.23
Rash	34 (11.8)				1 (0.4)				<0.001
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 1	Grade 2	Grade 3	Grade 4	
					number (percent)				
Other adverse events‡									
Infusion reactions	30 (10.4)	16 (5.6)	8 (2.8)	5 (1.7)	0	0	0	0	<0.001
Rash	114 (39.6)	107 (37.2)	34 (11.8)	0	32 (11.7)	11 (4.0)	1 (0.4)	0	<0.001
Hypomagnesemia§	95 (36.7)	28 (10.8)	7 (2.7)	8 (3.1)	29 (14.6)	1 (0.5)	0	0	<0.001

Con't

- Conversely, data that will not be used for analysis *should not be collected* on the CRF (or specified in protocol)
- If data collection is changed during the course of the study, the protocol should be amended

Design Principle #2:
Consider how data will be stored/analysed as CRFs being created

- If you will be creating a database to store/manipulate data abstracted on CRFs, do not make CRFs THEN consult database team or statistician.
- Design data collection/database **together**
- (Otherwise you will have wasted a lot of time….)

Design Principle #3 – Finalize CRFs Early

- Finalize CRFs prior to study start
- Often the process of creating CRFs will identify gaps, inconsistencies or errors in the protocol
- May require multiple iterations
- "Test" CRFs prior to rollout (for e.g. ask a CRA to see if they make sense)

Design Principle #4: Coding

- Collect "actual" data, rather than coded or interpreted information
- For example:
  - hematology/biochemistry value rather than a grade
  - dates rather than a calculated time interval
  - tumour measurements rather than only a response classification

Examples

NOT this ---	(answer)	But this ---	(answer
What is patient age?	58 years	What is date of birth?	Nov 24 1953 (age is actually **57**)
Worst grade of AST	Grade 3	AST UNL	298 (Grade **2**) 40
Best response	Stable disease	Tumour measures	Baseline: 20 mm Cycle 2: 25 mm (**Progressive Disease**)

Design principle #5:

Careful choice of categorical or continuous variables

- For categorical variables, consider/allow all possible values for each variable collected, including those that may be rare
  - For ovary histology: serous, clear cell, endometrioid, mucinous….
- Some variables may be continuous OR categorical: pros/cons to collecting each:

Example: Ovarian Cancer:
Residual disease after primary debulking

This?

- Please enter maximum size of residual disease:

  _ _._ cm

Or This?

- Maximum size of residual disease (check one)

□ None

□ Microscopic only

□ < 1 cm macroscopic

□ 1-5 cm

□ >5 cm

Design Principle #6:
Avoid Duplication

- Avoid capturing the same piece of data in more than one place on the CRF
- Inevitably values will differ and the error will take time and effort to correct

Design principle #7:
Avoid Free Text

- Avoid use of free text fields
- Free text is not analyzable without manual coding/interpretation
- Likewise, reduce use of vague variables such as "other" as a valid category

Free text not useful

Don't do this

- What adverse events did patient have (please type in):

Sore knees

Tiredness

Peeling soles of feet

Do this

- Select adverse events experienced from drop down list

Arthralgia

Fatigue

Palmar-plantar erythrodysesthesia

Design principle #8: Consider the user – especially if it will not be you!

- Provide instructions/definitions within the CRF to avoid misinterpretation, especially if there are multiple users
- Design with user in mind (e.g. order of data collection should flow appropriately)
- *Be unambiguous* (e.g. provide units expected)
- Be concise

Con't

- Avoid abbrs., unf. terms
  (Avoid abbreviations and unfamiliar terms)
- Be clear about format
  - 12h vs. 24h clock
  - European vs. US date order (YYYY MMM DD or DD MM YYYY)
- Full vs. partial dates permitted
- Permit "Not Done (ND)" or "unknown (UNK)" as options where appropriate

Case Report Forms to Database

- Paper CRF data are entered into a database/ database management system (eg Oracle, Access) that will allow sorting, calculation, analysis of information
- Electronic CRF data: create the database as they are entered
- For example:
  - data checking for accuracy, logic, missing values
  - ongoing monitoring of patient eligibility, safety, protocol compliance
  - analysis of large amounts of data

Output from database

- Need statistician/programmer help, here
- Be sure instructions/logic for calculation of output and sorting of variables is correct….
- Excel and other databases allow some analysis options as well

Your projects....

Phase I Radiosurgery + Sunitinib in Brain Metastases

Caroline Chung

Objectives

Primary Objective:

- Determine the safety and maximum tolerated dose of Sunitinib when combined concurrently with SRS in patients with 1-3 brain metastases

Secondary Objectives:

- To capture any observed late toxicities that may be attributable to this combined treatment of Sunitinib and SRS.
- Determine time to Intracranial Local Progression, and Intracranial Distant Progression
- Determine Brain Progression-free Survival
- Determine the influence of Sunitinib on the requirement for supportive corticosteroids.
- Quantify alterations in tumour perfusion parameters observed with dynamic contrast enhanced MRI (DCE-MRI) and DCE-CT
- Quantify normal tissue effects in brain tissue adjacent to metastatic lesions using MRI
- Assess serum biomarkers as potential prognostic or predictive factors
- To measure effect of SRS and sunitinib on neuropsychological function

From objectives to variables

Objective	What you will need to collect
Baseline info	Patient age (DOB), PS, underlying cancer ?prior therapy? Number/location brain mets Prior Rx for brain mets Neurological findings Imaging findings
Safety and MTD	Date registered on study Date start sunitinib Date of SRS Dose of sunitinib No. days of therapy Adverse events , grade and relation to therapy Stereotactic RS delivery
Late toxicities	Adverse events, grade, relationship post treatment until death
Time to Intracranial local and distant progression Brain PFS	Definition of local, distant progression in protocol Date of each Date of death Tumour measures to prove/confirm each

From objectives to variables

Objective	What you will need to collect
Influence of sunitinib on need for steroids	Concomitant steroid dose/dates (note: non-randomized trial so hard to interpret)
Quantify alterations in tumour perfusion with DCE-MRI and DCE CT	Need agreed methods /observer calculation of perfusion. Dates baseline/on study scans Perfusion parameters calculated for each time point
Quantify normal tissue effects	Need agreed methods /observer for normal effects Dates baseline/on study scans Normal effect parameters calculated for each time point
Serum biomarkers as potential prognostic/predictive factors	Identify each biomarker assay and output type (categorical/continuous) Database must contain data form baseline and on study values for each For prognostic: need all other variables entered into database likely to be prognostic For predictive: need control group?
Effect of SRS + sunitinib on neuropsychological function	Baseline and on study questionnaires Enter dates and values/answers for each question (have plan in place to handle missing data)

Genomic and proteomic assays subclassify TN breast tumour and predict outcome to chemotherapy

Maggie Cheung

Objectives

- Identify if genomic and proteomic signatures can help predict response to therapy in TN BC

(exploratory/discovery analysis; not a gene signature validation project)

From objectives to variables

Objective	What you will need to collect
Baseline clinical info	DOB PS HER2, ER, PR results ? Oncotype Dx results Date of diagnosis T size T location N M
Genomic data	At baseline and repeat sampling time points: Relative expression value for each gene tested for each patient
Therapy	Date/doses of each drug (neoadjuvant)
Outcome	Pathologic CR --date Clinical CR –date Clinical PR – date Etc.