LONGSCAN

- Data structure attributes & sample considerations

Terri Lewis, PhD
Coordinating Center
&
North Carolina Site

Administration Methods

- Face-to-face interview (F-to-F)– conducted when child participants were approximately 4, 6, 8, and 12.

- Annual Contact Interviews (ACI)– conducted on the years when face-to-face interviews were not (5, 7, 10, 11)

- Interviews were conducted separately with child participants and their caregivers

The Evolution of Data Collection Methods

- Ages 4 & 6 were interviewer administered: paper/pencil

- Age 8 was interviewer administered, but computer assisted

- Age 12 marked the beginning of the Audio-Enabled Computer Assisted Self-Administered (A-CASI) format.

Informants

- Child Informants – data collected at each F-to-F interview
- Caregiver (CG) Informants – data collected at each F-to-F interview and ACIs
- CPS – review cycles vary. Generally no more than 2 years pass without a comprehensive review of CPS records
- Teachers – collected at times corresponding to the F-to-F interviews assessing academic performance and school behavior, school safety, peer ratings, and engagement with school activities
- Interviewer –administered development and cognitive measures; provide ratings at each interview

Data Structure

Flat datasets: one observation per ID (wide)

Stacked datasets: multiple observations per ID (long)

Regardless of the informant, all data are linked by the (child) subject identification number (ID), which is combination of the study site and a unique numeric identifier

Data Structures

General rule of thumb:

- There is a dataset for each measure, administered to each respondent, at each time period.

Data Structures

- There are several ‘stacked’ datasets (e.g., CBCL, TRF, and CPS data). Multiple observations per ID due to:
  - Multiple visits (e.g., CBCL)
  - Multiple referrals (e.g., CPS data)
  - Multiple respondents of the same type (e.g., teachers- TRF)

Dataset Naming

- Dataset name can generally be broken down into components
  - Abbreviation of measure name
  - Form version
  - Data closure date

E.G., DEMA0404

Dem = demographics form

A = form version ‘A’

Retrieval date = April 2004

Dataset Naming

- Most dataset names are 8 characters with the 4 digit data closure date at the end.

- Exceptions:
    - data from the age 8-11 Data Management System (DMS). These datasets will have a 3 character form name + version followed by the 4-digit retrieval date (DEA0708)
    - Datasets we have combined

      (MRCA + MRCB = MRC0404)

Dataset Naming

- Some measures were administered at multiple time points, but are housed in different datasets, due primarily to the version of the DMS in use at the time or to slight variations in wording, response options, question order, etc.

Dataset Naming

- The name of the datasets for these measures will retain the primary mnemonic, but with a different ‘form version’ at the end.

E.G., The datasets containing caregiver demographics are

DEMA (age 4)

DE6A (age 6)

DEA (age 8)

DEMB (age 12)

Dataset Naming

- If the same construct is assessed, but is assessed with different measures over time, the mnemonic generally represents the measure name, not the construct.

- Example, caregiver depression datasets include:
- DEPA (CES-D; ages 4 & 6)
- BSA (BSI; age 8)
- DEPB (CES-D; age 12)

Dataset Naming

- Some of the measures have been scored. The scored data are housed in their own datasets, separate from the item-level data. Generally, scored datasets have the partial mnemonic of the item level dataset, followed by an ‘s’

E.G., the item level data from the Child Behavior Checklist is located in the CBCL dataset. The scores for the CBCL are located in the CBCS dataset.

Constructing Analysis Datasets

- Because the LONGSCAN datasets are a combination of flat and stacked data structures, and due to the volume of datasets, it is NOT recommended that users attempt to merge all of the LONGSCAN datasets together into one dataset.

Constructing Analysis Datasets

- It is recommended that users determine, based on their analysis questions, what datasets and/or variables from those datasets are most relevant. Once these are selected and the structure of the datasets are compatible, then a 1:1 merge can be done, linking by ID.

- Alternatively, if the analysis technique requires a stacked dataset, then similar steps should be taken to ensure correct merging of the data (i.e., subject id & visit number)

How to Link Observations

- All datasets/observations can be linked by the subject ID:

Var name = ID

- For stacked datasets, observations can be linked by subject ID and Visit Number

ID = Subject ID

Visit = Visit Number

Documentation & Sources of Information

- There are two critical components to understanding and identifying the data of interest. These components should always be used in conjunction with each other as each has unique information to offer.

- LONGSCAN Measures Manuals
  - Volumes 1, 2, & 3.
    - http://www.iprc.unc.edu/longscan/pages/measures/index.htm

- Data Dictionaries

LONGSCAN Measures Manuals

- Three Measures Manuals, corresponding to the three developmental periods

  early childhood

  middle-childhood

  early adolescence

Measures Manuals

Description of Measure

- Purpose
- Conceptual Organization
- Item Origin/Selection Process
- Materials
- Administration Method
- Training

Measures Manuals

- Scoring
  - Score types
  - Score Interpretation
- Norms and/or Comparative Data
- LONGSCAN Use
  - Data Points
  - Respondent
  - Mnemonic & Version
  - Rationale
  - Administration & Scoring Notes

Measures Manuals

- Results
  - Descriptive Statistics
  - Reliability & Validity
- References & Bibliography

LONGSCAN Data Dictionaries

- The Data Dictionaries (DD) provide detailed information on the items and response options that exist in the dataset.

- The scored data for any given measure exists as it's own dataset and is not included with the item level data.

- Each DD for scored data include the algorithms for the derivation of the scores and any other general information on use or interpretation.

LONGSCAN Data Dictionaries

- The DDs are arranged in the following order:

  Table of contents

  DDs for item-level data

  DDs for scored data

  Appendices that are relevant to the datasets

Adolescent Delinquency Survey – ADSA

Variable Name	Format	Variable Description	Coding if Categorical
ID	Char	Longscan Subject ID	

Center	Char	Longscan Field Center	EA = East
			MW = Midwest
			SO = South
			SW = Southwest

Visit	Num	Visit Number	EA = 12
			MW = 12
			SO = 12
			SW = 12
			NW = 12

Adolescent Delinquency Survey – ADSA

Variable Name	Format	Variable Description	Coding if Categorical
ADSA1	NUM	Did you ever take part in gang activities	0 = NO 1 = YES
ADSA2	NUM	Did you belong to a group that other people consider a gang	0 = NO 1 = YES
ADSA3	Num	Did you steal or shoplift	0 = NO 1 = YES
ADSA4	Num	Were you in a physical fight	0 = NO 1 = YES
ADSA4A	Num	How many times were you in a physical fight?	1 = 1 time 2 = 2-5 times 3 = 6-12 times 4 = > 12 times

Maltreatment Data

Data Types

CPS Case Record Reviews (RNAB0708)

Data derived from the RNAB dataset (M_SD0810)

Caregiver Report of Sexual Abuse
Child Sexual Behavior (SBA – age 8)
Sexual Abuse of Child (SAC – age 12)

Youth Self-Report at Age 12 –
PHYA (PHYS) – Physical Abuse
PSMA (PSMS) – Psychological Abuse
SARA (SARS)   – Sexual Abuse (Supplement Data = SASA)
AMPA   - Neglect

Maltreatment Data

Other Possible Data Types

Caregiver
Conflict Tactics Scale: Parent to Child
CTSB (CTSS) Ages 4* & 6
DMA (DMS) Age 8
PCCT (PCCS) Age 12

SW site did not administer at age 4
The NW site modified the response options due to IRB concerns – be sure to consult Measures' Manual for details and site variations in administration.

CPS Case Record Reviews

- Case records from CPS are reviewed for each subject (with current consent).
- Tri-coding of allegations & findings
  - CPS labels of maltreatment allegations, findings, & risk factors
  - NIS coding of allegation and findings' narratives
  - MMCS coding of allegation and findings' narratives

- Each type of coding offers different perspective and different information

RNAB

- Observations in the CPS data reflect a referral to CPS
  - Up to 6 allegations of maltreatment may be coded for any given referral
  - The number of observations (referrals) will vary across LONGSCAN participants
  - The data are not organized in the dataset by age of the participant or interview cycle (visit number is of no use)

RNAB Data

- CPS data are structured to provide the most flexible use of the data, but may require a considerable amount of work depending on the questions of interest and dataset structure necessary for analyses

- A tutorial on the use of these data accompany the data dictionary for the RNAB

Example of RNAB Structure

ID	Review Date	Ref Date	Incdnt Date	First Maltx Code	Sverity	Perp #1	Gender of Perp #1	Second Maltx Code	Perp #1	Gender of Perp #1
01	3/5/07	5/4/05	5/2/05	200	2	5	2	403	1	1
02	2/9/95	7/4/93	6/30/93	304	1	1	1			
02	2/9/95	7/4/93	12/1/93	103	2	1	1	500	1	1
02	6/30/97	3/5/95	3/1/95	401	1	1	1			
03	5/10/93	6/15/92	1/1/92	105	4	1	2			

Structure of RNAB Dataset

Number of Records per Subject	Count of Subjects	Count of Observations in Dataset
1	214	214
2	154	522
3	133	921
4	95	1301
5	69	1646
6	53	1964
7	37	2223
8	35	2503
9	30	2773
10	19	2963
11	14	3117
12	14	3285
13	12	3441
14	8	3553
15	9	3688
16	6	3784
17	3	3835
18	6	3943
19	1	3962
20	2	4002
Total	914	4002

440 participants do not have a records in the RNAB datasets.

Alternative (Maltreatment) Dataset: M_SD0708

- The M_SD dataset was developed to make the CPS data easier to work with. The M_SD contains 1 observation for each LONGSCAN participant. The variables are relevant RNAB data, aggregated to correspond to the F-to-F interviews.

M_SD0810 Dataset

Included are:
- allegations
- referrals
- CPS determinents & referrals for domestic violence, based on referral/summary narrative information

Classification of maltreatment coded by:
- type [physical, sexual, emotional abuse, neglect (FTP, LOS, EDU),
  moral/legal, drugs/alcohol]
- single/multiple
- combinations of maltx types (expanded hiearchical type)
- severity
- chronicity of maltreatment from birth to age 9.5.

Available for time frames 0-4, 4-6, 6-8, 8-10, 10-12, and 8-12.

Example of M_SD Structure

ID	Center	# Phy Abuse Alleg 0-4	# Phy Abuse Subst 0-4	Max Sevrty Phy Abuse 0-4	Single Vs Mltpl Type 0-4	# Phy Abuse Alleg 4-6	# Phy Abuse Subst 4-6	Max Sevrty Phy Abuse 4-6	Single Vs Mltpl Type 4-6
01	SW	2	0	1	1	0	0	0	.
02	NW	5	4	2	1	0	0	0	.
03	MW	0	0	0	0	0	0	0	.
04	EA	0	0	.	.	0	0	.	.

M_SD Derived Data

Because the M_SD represents aggregate data, data specific to any given allegation/substantiation/ referral is not available (e.g, perpetrator data, severity for a given allegation of maltreatment, etc.).

If an individual is interested in a time frame or specific age not included in the M_SD, the user would need to work with the RNAB data.

Notes of Caution

(1) The absence of an observation in the RNAB, does not necessarily, mean that there was no maltreatment.

(2) The aggregated data in the M_SD does assume that no observation = no maltreatment and/or no maltreatment of that type, and is dependent on the RNAB data and thus subject to any issues inherent in the collection of CPS data

Datasets with Useful Variables

- IDS_0708 (flat)
    - Child gender
    - Child race
    - *Child date of birth
    - Interview indicators (child & caregiver)
    - Study site

Datasets with Useful Variables

- Cover Sheets (flat)
  - Caregiver Respondent Relationship to child participant
  - Date of the interview

Child: CRC (4)* CICA (6) CIA (8) CICB (12)

CG : MRC (4) PRCA (6) PIA (8) PRCB (12)

ACI : ACIA (5, 7) ACA (9) ACB (10, 11)

Datasets with Useful Variables

Derived Household Composition (DHC0810)

- Dataset of derived variables aggregated over the measures of household composition from ages 4, 6, 8, and 12.

Includes

- respondent's gender & relationship to child
- foster status of caregiver
- # of adults, children, & total in household
- Indicators for presence of household members (e.g., bio mom, grandmother, non-relative female, etc.)
- multigenerational households
- basic family composition types
- basic family structure types
- living arrangements

Caregiver Arrangement

- How to determine ‘foster care’

  (1) All participants from the SW site were removed prior to age 4 and placed in foster care.

  (2) The caregiver relationship to the youth at the time of the interview (& ACI)

  (3) Household composition

  (4) Life Events Scale for Children

Caregiver Arrangement

- Respondent Relationship to Child & Household Composition datasets
  - Cover sheets from F-to-F and ACI Interviews

    At ages 4 & 6 did not make the distinction between kin & non-kin foster caregivers
  - Household Composition Forms: HOMA (4,6) FCA (8), FCHB (12)
  - Derived Household Composition Dataset (DHC0810)

Caregiver Arrangement

- Life Events Datasets

…in the last year, has child moved away from family,

# times moved into foster care (or group home/shelter)

LECA (5*, 6, 7)

LEB (8, 9, 10, 11)

LECC (12)

At ACI 5, BA did not administer.
Note the response option for the LECC is YES/NO, not the # of times.

Caregiver Respondents (%)

	Age 4	Age 6	Age 8	Age 12
Bio Mom	72	70	68	64
Grandmother	7	9	8	8
Foster Mother*	6	4	2	2
Adoptive Mother	4	7	9	10
Other Female Relative	4	0	2	4
Biological Father	3	3	3	5
Other Female	2	5	4	0.6

At ages 8 & 12, the endorsement for foster mother is only ‘kinship foster mother’

Caregiver History of Loss & Victimization

- Caregiver Loss & Victimization was assessed at Age 4. Assessment was split into two measures:

  Loss: LSSA

  Victimization: VICA

The SW site did NOT administer the VICA.

Data Considerations

- LONGSCAN defines the baseline sample as those completing either an age 4 or age 6 interview. There are 104 participants with an age 6, but no age 4 interview.

Data Considerations

- Sites samples vary by maltreatment risk and entrance into the LONGSCAN Consortium. Oldest participants are from the Southern site, the youngest from the Midwest Site.

- Age distribution within sites vary with the exception of the Southern Site.

Start and End Dates of Data Collection by Interview

	Date of first interview	Date of last interview
Age 4	7/25/91	3/20/00
Age 6	3/2/93	2/8/02
Age 8	12/20/94	7/17/03
Age 12	8/25/98	10/6/07

Age Range of Sample by Interview

			Range	
Interview	Mean	Std	Min	Max
4	4.5	0.7	3.5	7.5
6	6.2	0.5	5.1	9.0
8	8.3	0.4	6.6	10.2
12	12.4	0.4	10.4	14.2

Attrition in LONGSCAN
Types of Attrition

- Approached – but did not consent
- Consented – but did not participate
- Participated – but did not complete

Types of Attrition

- Approached – but did not consent
  - limited data – not cross-site
- Consented – but did not participate
- Participated – but did not complete

Types of Attrition

- Approached – but did not consent
  - limited data – not cross-site
- Consented – but did not participate
  - completed baseline interview only
- Participated – but did not complete

Types of Attrition

- Approached – but did not consent
  - limited data – not cross-site
- Consented – but did not participate
  - completed baseline interview only
- Participated – but did not complete
  - # of completed interviews vary across individuals
  - sequence of responses varies across individuals

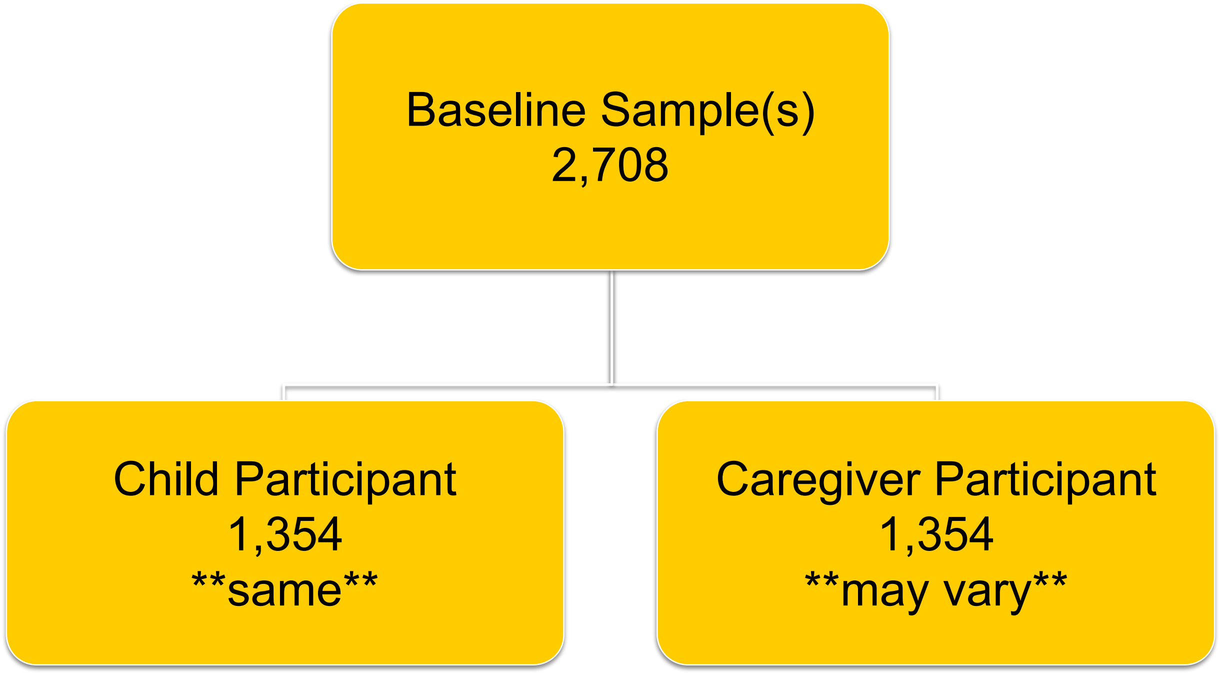
Reasons for Attrition

- Death
- Participant Withdrawal
- Lack of Contact

Types of Missing Data

- Item Non-Response
  - MCAR
  - MAR
  - NMAR
- Unit Non-Response
  - CRD (completely at random)
  - RD (random dropout)
  - ID – (informative dropout)

Starting Sample



Baseline Sample(s)
2,708

Child Participant
1,354
same

Caregiver Participant
1,354
may vary

Interview Completion

- Child Interview = X
- Caregiver Interview = X
- Child OR Caregiver Interview = X
- Child AND Caregiver Interview = X

Conceptualizing Attrition in LS

Number of interviews completed

- Issue 1 – those added at T6 will have fewer interviews than those starting at T4

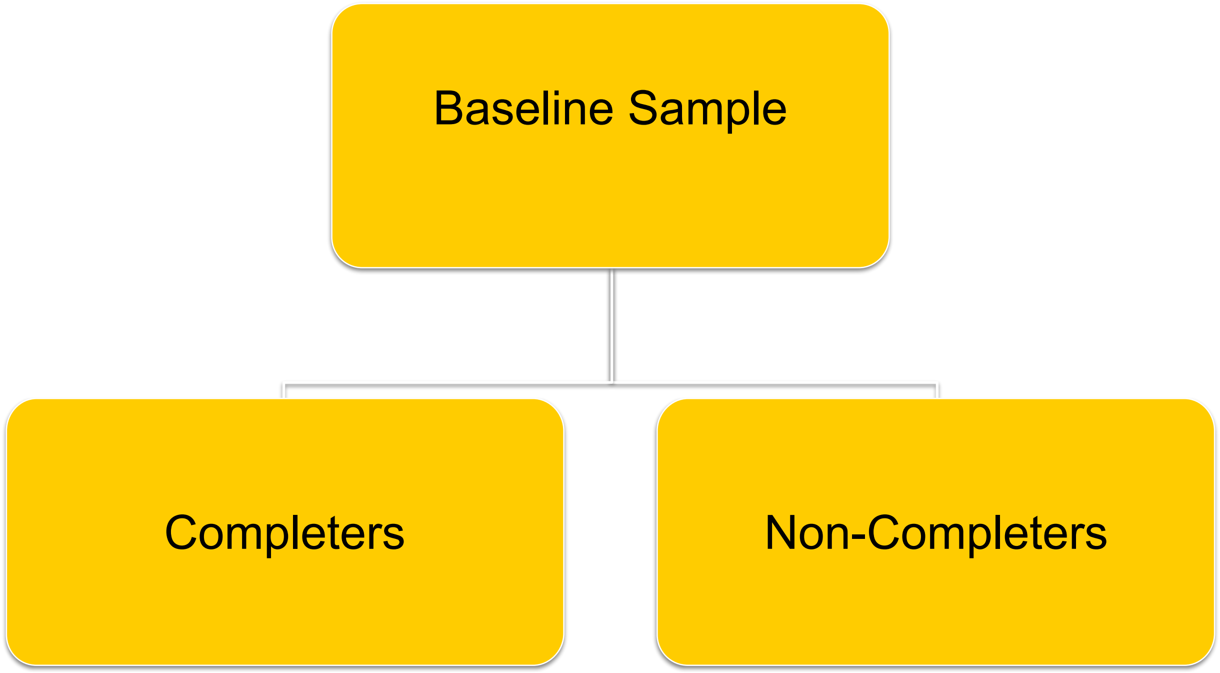
Conceptualizing Attrition in LS

Number of interviews completed

- Issue 1 – those added at T6 will have fewer interviews than those starting at T4

- Issue 2 – those still ‘active’ are not ‘attritted’ even if they have not completed the full sequence of interviews

Conceptualizing Attrition in LS

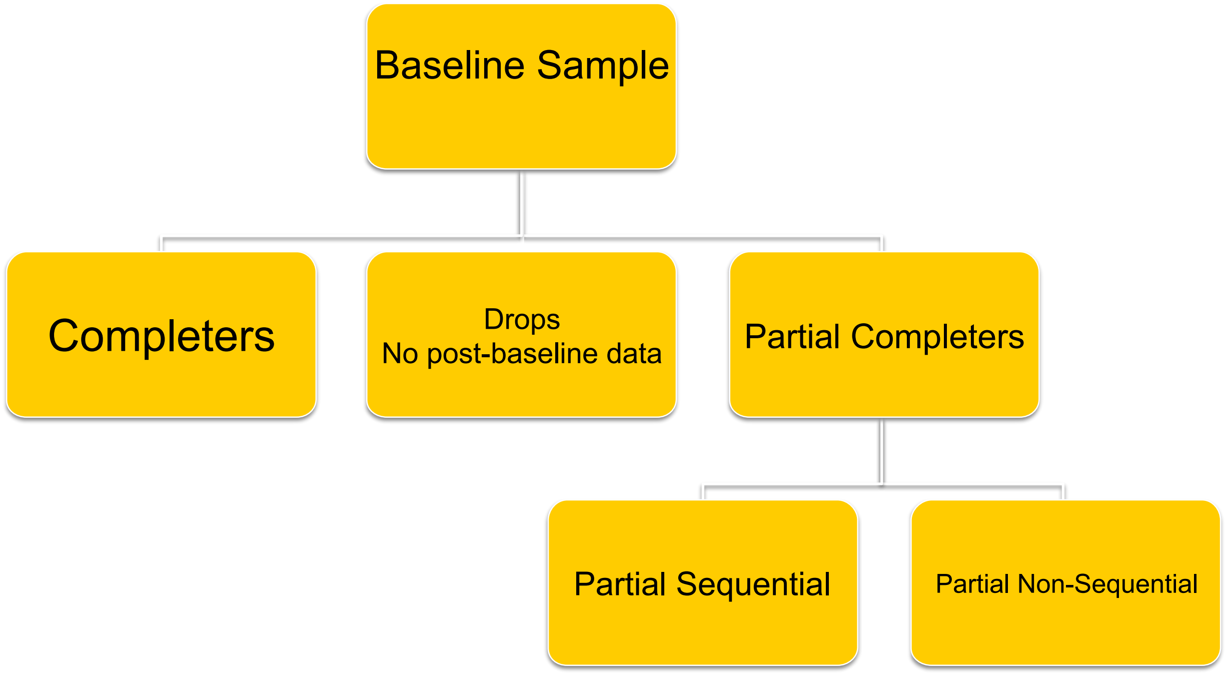


Baseline Sample

Completers

Non-Completers

Conceptualizing Attrition in LS



Baseline Sample

Completers

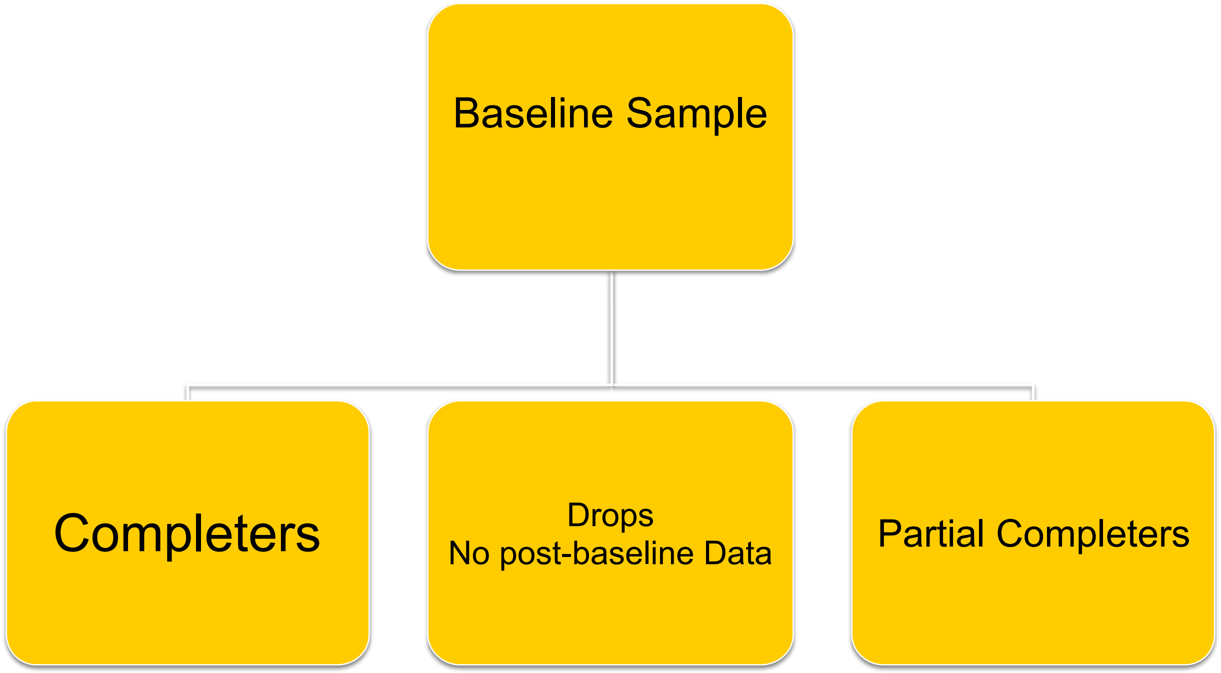
Drops
No post-baseline data

Partial Completers

Partial Sequential

Partial Non-Sequential

Conceptualizing Attrition in LS



Baseline Sample

Completers

Drops
No post-baseline Data

Partial Completers

Interview Completion Rate

Completion Rate	Age 4	Age 6	Age 8	Age 12
	92% (1250)	91% (1236)	84% (1140)	72% (976)

# Interviews Completed	4	3	2	1

M = 3.40,
SD = .86
	60%
	(810)
		26%
		(347)
			9%
			(124)
				5%
				(73)

Retention Rate

Retention	Age 4 - 6	Age 6-8	Age 8-12
	84% (1132)	81% (1093)	67% (910)

Percentage of Completion Category: Baseline-T12

	Completers	Partial Completers	Suspected Drops
%	65	30	5
N	878	403	73

Suspected Drops have completed 1 interview only.

Partial completers have completed more than 1 but fewer than 4.

Completers have completed all four interviews.

Distribution of Sample Attributes by Completion Group

	Completers (n=878)	Partial Completers (n = 403)	Suspected Drops (n = 73)
Gender			
Female	51	53	56
Male	49	47	44
Race			
African American	55	53	42
White	26	25	36
Mixed Race	11	12	15
Other	8	11	7

Distribution of Sample Attributes by Completion Group

	Completers	Partial Completers	Suspected Drops
Status at Recruitment			
Reported	61	60	55
High-Risk	22	24	29
Control	17	16	16
Maltreated by Age 4	57	56	60
Site			
East	19	24	23
Midwest	19	15	18
South	18	16	26
Southwest	23	28	16
Northwest	20	16	16

Partial Completers
Sequence of Interview Completions

T4	T6	T8	T12	n	%
1	1	1	0	198	49
1	1	0	0	75	19
1	1	0	1	49	12
1	0	1	1	32	8
0	1	1	0	17	4
1	0	1	0	15	4
1	0	0	1	12	3
0	1	0	1	5	1

Methods to deal with missing data (Abraham & Russell, 2004)

- Ad hoc Methods
  - complete case/available case
- Single imputation Methods
  - E.G., LOCF
- Model-based Methods
  - GEE, MLE, FIML
- Multiple Imputation
- MAR
  - Selection Models
  - Pattern-Mixture Models

References

- Abraham, W.T., & Russell, D.W. (2004). Missing data: A review of current methods and applications in epidemiologic research. Current Opinion in Psychiatry, 17, 315-321.
- Ahern, K., & Le Brocque, R. (2005). Methodological issues in the effects of attrition: Simple solutions for social scientists. Field Methods, 17, 53-69.
- Goodman, J. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. Journal of Management.
- Mazumdar, S., Tang, G., Houck, P.R., Dew, M.A, et al., (2006). Statistical analysis of longitudinal psychiatric data with dropouts. Journal of Psychiatric Research, 41 1032-1041.