Preventive Medicine Grant Writing Seminar Series

Session 7: Data Collection, Data Management, Storage & Quality Assurance

Phyllis A. Richey, Ph.D.
Professor, Departments of Preventive Medicine, Physical Therapy and Pediatrics
Clinical & Community Research Informatics Section (CCRIS)

2

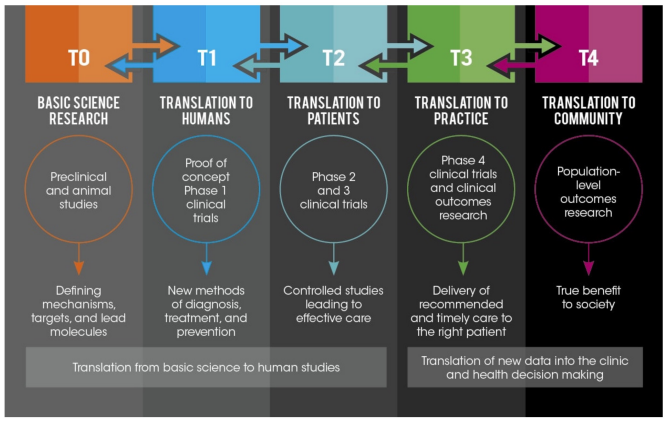
Overview

- **Browner et al chapters 18-19**
- **Informatics vs Research Informatics**
- **Data Collection, Management Defined Entry Systems**
  - Database design for study implementation, participant recruitment and retention, outcome measure data collection, intervention delivery
  - Data standardization
- **Technology Resources**
  - Where used
  - How processed
- **Programming, Data Entry, Cleaning, Analysis & Security**
- **Special insights on research data management design**

3

Informatics vs. Research Informatics

- **INFORMATICS**
  - The science of processing data for storage and retrieval.
- **RESEARCH INFORMATICS**
  - Sub-discipline within biomedical informatics which focuses on developing new theories, tools, and solutions used to translate data across the basic research, clinical trials, medical center and community practice continuum.



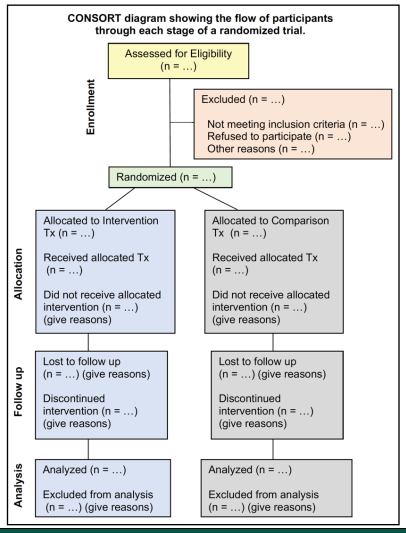
*National Center for Advancing Translational Sciences (NCATS), 2023

Why is a research informatics solution critical to successfully conducting a research study?

- **It defines the procedures for information (data) processing and storage during the study**
  - Determines the methods of data collection and storage
  - Establishes user interface (GUI) through which information is entered, navigated, and presented in a meaningful way.
  - Provides manipulation of data and creation of new information through rules, calculations, index lookups

Development of a research informatics solution:

- Defines “HOW” the study is to be conducted
  - Participant Recruitment and Enrollment - Where the CONSORT begins
  - Eligibility and Randomization
  - Intervention Delivery
  - Data Collection – outcome and process measures
  - Storage & Mining of Study Information
  - Data Sharing (central repository)
  - Ongoing Quality Assurance for Data Integrity
  - Data Security Maintenance – throughout the life of the project



Research Informatics Solutions

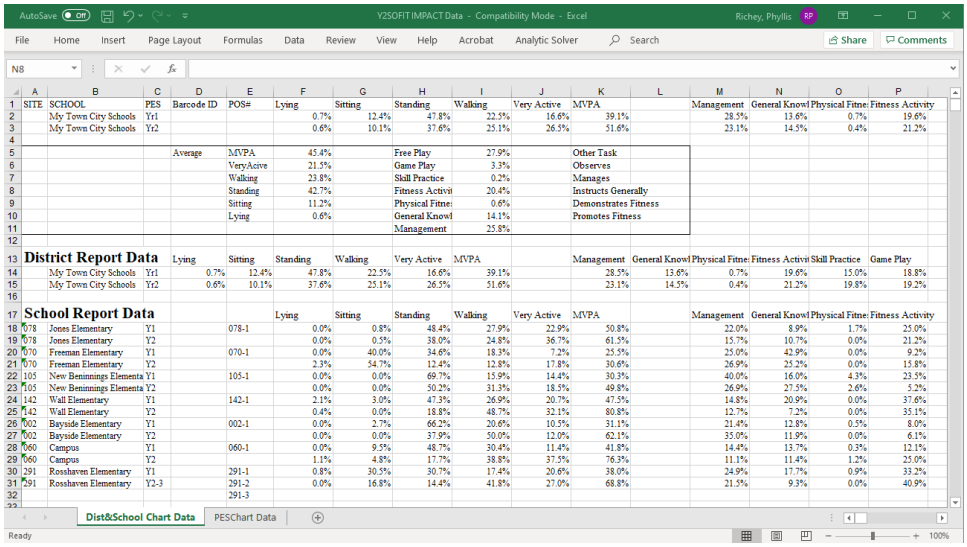
- Common Types of Data Management Resources
  - Single spreadsheet
  - Statistical
  - Database Management System (DMBS)
- The Foundational Unit – The Data Table

So … What's a data table?

- **Organized representation of data into columns and rows**
- **Data organization is strictly restricted**
  - One **variable** (field) per **column**
  - One **record** per **row**

8

Example of what is NOT a Data Table



SITE	SCHOOL	PES	Barcode ID	POS#	Lying	Sitting	Standing	Walking	Very Active	MVPA		Management	General Knowl	Physical Fitne	Fitness Activity
	My Town City Schools	Yr1			0.7%	12.4%	47.8%	22.5%	16.6%	39.1%		28.5%	13.6%	0.7%	19.6%
	My Town City Schools	Yr2			0.6%	10.1%	37.6%	25.1%	26.5%	51.6%		23.1%	14.5%	0.4%	21.2%

	Average	MVPA	45.4%		Free Play	27.9%		Other Task
		VeryAcive	21.5%		Game Play	3.3%		Observes
		Walking	23.8%		Skill Practice	0.2%		Manages
		Standing	42.7%		Fitness Activit	20.4%		Instructs Generally
		Sitting	11.2%		Physical Fitne	0.6%		Demonstrates Fitness
		Lying	0.6%		General Knowl	14.1%		Promotes Fitness
					Management	25.8%		

District Report Data

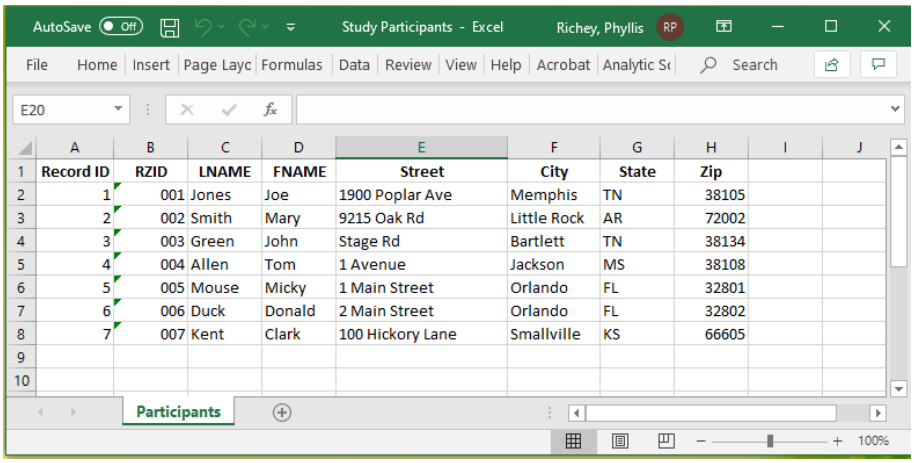
		Lying	Sitting	Standing	Walking	Very Active	MVPA		Management	General Knowl	Physical Fitne	Fitness Activit	Skill Practice	Game Play
	My Town City Schools	Yr1	0.7%	12.4%	47.8%	22.5%	16.6%	39.1%	28.5%	13.6%	0.7%	19.6%	15.0%	18.8%
	My Town City Schools	Yr2	0.6%	10.1%	37.6%	25.1%	26.5%	51.6%	23.1%	14.5%	0.4%	21.2%	19.8%	19.2%

School Report Data

				Lying	Sitting	Standing	Walking	Very Active	MVPA	Management	General Knowl	Physical Fitne	Fitness Activity
078	Jones Elementary	Y1	078-1	0.0%	0.8%	48.4%	27.9%	22.9%	50.8%	22.0%	8.9%	1.7%	25.0%
078	Jones Elementary	Y2		0.0%	0.5%	38.0%	24.8%	36.7%	61.5%	15.7%	10.7%	0.0%	21.2%
070	Freeman Elementary	Y1	070-1	0.0%	40.0%	34.6%	18.3%	7.2%	25.5%	25.0%	42.9%	0.0%	9.2%
070	Freeman Elementary	Y2		2.3%	54.7%	12.4%	12.8%	17.8%	30.6%	26.9%	25.2%	0.0%	15.8%
105	New Beninnings Elementa	Y1	105-1	0.0%	0.0%	69.7%	15.9%	14.4%	30.3%	40.0%	16.0%	4.3%	23.5%
105	New Beninnings Elementa	Y2		0.0%	0.0%	50.2%	31.3%	18.5%	49.8%	26.9%	27.5%	2.6%	5.2%
142	Wall Elementary	Y1	142-1	2.1%	3.0%	47.3%	26.9%	20.7%	47.5%	14.8%	20.9%	0.0%	37.6%
142	Wall Elementary	Y2		0.4%	0.0%	18.8%	48.7%	32.1%	80.8%	12.7%	7.2%	0.0%	35.1%
002	Bayside Elementary	Y1	002-1	0.0%	2.7%	66.2%	20.6%	10.5%	31.1%	21.4%	12.8%	0.5%	8.0%
002	Bayside Elementary	Y2		0.0%	0.0%	37.9%	50.0%	12.0%	62.1%	35.0%	11.9%	0.0%	6.1%
060	Campus	Y1	060-1	0.0%	9.5%	48.7%	30.4%	11.4%	41.8%	14.4%	13.7%	0.3%	12.1%
060	Campus	Y2		1.1%	4.8%	17.7%	38.8%	37.5%	76.3%	11.1%	11.4%	1.2%	25.0%
291	Rosshaven Elementary	Y1	291-1	0.8%	30.5%	30.7%	17.4%	20.6%	38.0%	24.9%	17.7%	0.9%	33.2%
291	Rosshaven Elementary	Y2-3	291-2	0.0%	16.8%	14.4%	41.8%	27.0%	68.8%	21.5%	9.3%	0.0%	40.9%
			291-3										

9

Example of a TRUE Data Table



Record ID	RZID	LNAME	FNAME	Street	City	State	Zip
1	001	Jones	Joe	1900 Poplar Ave	Memphis	TN	38105
2	002	Smith	Mary	9215 Oak Rd	Little Rock	AR	72002
3	003	Green	John	Stage Rd	Bartlett	TN	38134
4	004	Allen	Tom	1 Avenue	Jackson	MS	38108
5	005	Mouse	Micky	1 Main Street	Orlando	FL	32801
6	006	Duck	Donald	2 Main Street	Orlando	FL	32802
7	007	Kent	Clark	100 Hickory Lane	Smallville	KS	66605

Data Management Resource Types

- **Single spreadsheet**
  - Most basic - Limited to two-dimensional data table format
- **Relational Database**
  - Allows for more complex data format
- **Each data management resource type's unique features = its**

"Data Management Model"

ParticipantID	FName	DOB	Sex	Hyperbili_ind	ExDate	ExWght	ExHght	IQ
2101	Robert	1/6/2010	M	1	1/29/2015	23.9	118	104
2322	Helen	1/6/2010	F	0	1/29/2015	18.3	109	94
2376	Amy	1/13/2010	F	1	3/22/2015	18.5	117	85
2390	Alejandro	1/14/2010	M	0				
2497	Isiah	1/18/2010	M	0	2/18/2015	20.5	121	74
2569	Joshua	1/23/2010	M	1	2/13/2015	24.8	113	115
2819	Ryan	1/26/2010	M	0				
3019	Morgan	1/29/2010	F	0	2/9/2015	19.1	105	105
3031	Cody	2/15/2010	M	0	4/16/2015	15.2	107	132
3290	Amy	2/16/2010	F	1	4/12/2015	18.0	102	125
3374	Zachary	2/21/2010	M	1				
3625	David	2/22/2010	M	1	2/10/2015	19.2	114	134
3901	Jackson	2/28/2010	M	0				

FIGURE 19.1 Simplified data table in "datasheet view" for a cohort study of the association between neonatal hyperbilirubinemia and IQ score at age 5. The binary predictor is "Hyperbili_ind," defined as whether the total bilirubin rose to 25 mg/dL or more in the first 10 days after birth, and the continuous outcome is "IQ," the participant's IQ score at age 5. Participants 2390, 2819, 3374, and 3901 were not examined at age 5.

FIGURE 19.2 A two-table infant jaundice study database with a table of study participants (in which each row corresponds to a single study participant) and a table of examinations (in which each row corresponds to a particular examination). For example, Participant 2322 is identified as Helen, date of birth 1/6/2010, in the first table, and has three exams in the anonymous second table. Note that ExWght and ExHght are entered in the exam table, not the participant table.

(*Browner* et al, Designing Clinical Research 5th Edition, 2023)

Advantages of Database Management System (DBMS) Modeling

- **What is Data Modeling?**
  - Techniques, not rules
  - Real world problem translated to a digital structure
  - Good modeling is essential

Types of DBMS Models

- **Depends on how the data is organized**
- **Flat Design Model**
  - Presents data in a single table or list (address list)
  - Flat design models have a “One to One Relationship”
  - Fields represent all parameters (prone to duplicate data)
  - Functionality limited to storing information, manipulate fields, printing and displaying data

	A	B	C	D	E	F	G	H
1	Record ID	RZID	LNAME	FNAME	Street	City	State	Zip
2	1	001	Jones	Joe	1900 Poplar Ave	Memphis	TN	38105
3	2	002	Smith	Mary	9215 Oak Rd	Little Rock	AR	72002
4	3	003	Green	John	Stage Rd	Bartlett	TN	38134
5	4	004	Allen	Tom	1 Avenue	Jackson	MS	38108
6	5	005	Mouse	Micky	1 Main Street	Orlando	FL	32801
7	6	006	Duck	Donald	2 Main Street	Orlando	FL	32802
8	7	007	Kent	Clark	100 Hickory Lane	Smallville	KS	66605

Types of DBMS Models

- **Relational Design Model**
  - Developed by E.F. Codd in 1970 (Ref: “A Relational Model of Data for large Shared Data Banks”)
  - Excellent for data entry
  - Utilizes multiple (flat) tables - single data points per field (normalized)
  - Eliminates redundancy through “normalization” (smoothes out data clumps)
  - Functionality to perform queries, join tables, establish integrity constraints, create reports
  - Utilizes unique fields defined as indexes (keys) establishing connections (relationships) between tables
  - Useful in sharing data across networks, internet, electronic devices, other (software) programs
  - Relational Models can have One to One, One to Many, Many to One or Many to Many Relationships

One versus Many

Defining Different Types of Relationships

- **One to one (typical flat table)**
  - Person and Social Security Number
  - Patient and blood type
  - Rare to see these in separate tables

	A	B	C
1	**First Name**	**Last Name**	**Blood Type**
2	Mary	Green	A+
3	John	Smith	A+
4	Glenda	Robbinson	O+
5	Susan	Jones	B-
6	Bobby	Taylor	A-
7	Brian	Williams	B+
8	Fredrick	Miller	O-
9	Alice	Washington	B+

One versus Many
Defining Different Types of Relationships

- **One to many / many to one**
  - Team to player
  - Blood samples to patient

Blood Samples

Sample ID	Patient ID	Date	Time
1	003	10/25/21	6:00
2	004	1/5/22	12:16
3	001	12/5/19	13:56
4	002	1/2/20	20:03
5	005	11/16/20	6:30
6	008	8/8/22	6:42
7	001	4/3/22	16:01
8	006	11/1/21	9:14
9	003	2/19/21	23:00
10	004	7/7/19	18:37
11	006	11/4/20	3:15
12	007	3/5/21	14:59
13	008	3/5/21	10:45
14	009	2/5/22	11:07
15	010	9/5/21	12:45
16	006	12/15/19	12:05
17	004	4/5/22	14:50
18	005	6/18/20	15:19
19	006	1/25/22	16:45
20	007	7/5/19	17:49

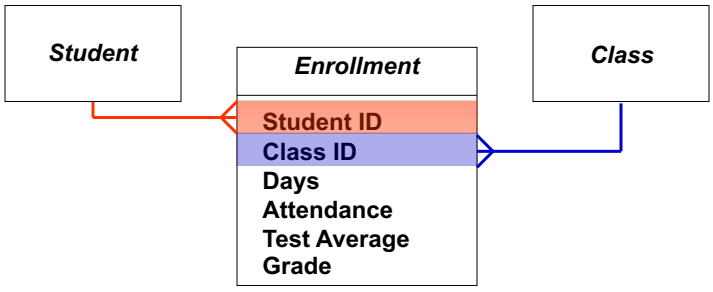
Patients

Patient ID	First Name	Last Name	Blood Type
001	Mary	Green	A+
002	John	Smith	A+
003	Glenda	Robbinson	O+
004	Susan	Jones	B-
005	Bobby	Taylor	A-
006	Brian	Williams	B+
007	Fredrick	Miller	O-
008	Alice	Washington	B+
009	Kelly	Franklin	A+
010	Bill	Davis	O+

(Browner et al, Designing Clinical Research 5th Edition, 2023)

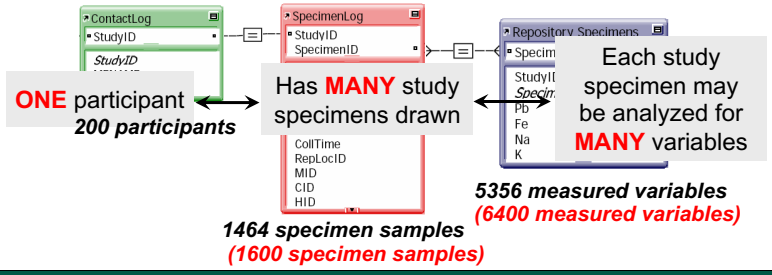
Modeling: Many to Many Relationships

- **How do we model these relationships?**
- **Two one to many relationships**
- **Can't just use one key (connector) between these 2 files**
- **We need a third table "in the middle"**



Many to Many Relationships
Role in Efficient Data Storage

- **The most complex relationship type**
  - Example:
    - Patient blood samples measured for concurrent medications
- **Generally, need to do extra work to resolve these**



Common Methods of Data Collection & Storage

Software Programs Used in Database Design, Data Management & Analysis

Spreadsheet	Integrated Desktop Relational Database	Enterprise Database	Integrated Web-Based Data Mgmt	Machine Readable Forms & Online Survey	Statistical Analysis
MS Excel	FileMaker Desktop, Server, App* (Cross platform)	Oracle	REDCap*	Qualtrics	R*
Google Drive Spreadsheet (Sheets*)	MS Access Office 365 - Desktop (PC only)	SQL	MediData RAVE	QuestionPro	SAS
Apache OpenOffice*		MySQL	EpiInfo*	TeleForm	SPSS
		PostgreSQL*	Datalabs EDC	Zoomerang	Stata
			QuesGen	SurveyMonkey	JMP
			OnCore		

* *Free*

(Partially reproduced from Browner et al, Designing Clinical Research 5th Edition, 2023, Table 19.1)

DBMS Programming Components

- **Data Tables**
- **Data Modeling Relationships**
  - Primary and Foreign Keys
  - Normalization
- **Data Dictionaries (Metadata)**
  - Field name (coding conventions)
  - Field data type (text, number, date, time, container)
  - Field definition
- **Field validation constraints**
  - Auto-entered, calculated, looked-up values
  - Restricted data type, not empty/unique, within range, value list
- **Data entry control style**
  - Edit box, value pick list, radio button, check box, calendar

DBMS Data Entry Methods

- **Case report form (paper form with manual transcription)**
- **OCR (Optical Character Recognition) - TeleForm**
- **Direct electronic data capture**
  - **Internal DBMS layouts – desktop, laptop, mobile device (FileMaker Go App)**
- **Indirect data capture**
  - **External data import protocols – Open Database Connectivity (ODBC), Java Database Connectivity (JDBC)**
    - **Survey tools – Qualtrics, QuestionPro, REDCap, SurveyMonkey**
    - **Technology devices – Accelerometers, BP/Holter monitors, body weight scales, lab equipment**
    - **Apps – LoseIt, MyFitnessPal**
    - **Online Validated Instruments – NCI's Diet History Questionnaire (DHQ)**

DBMS Data Cleaning, Analysis and Confidentiality

- **Queries and Reports**
  - Filter data using queries
  - Check for errors
  - Summarize data in reports
- **Statistical Analysis**
  - Extract “cleaned” datasets
- **De-identification & Resource Sharing**

DBMS Security

- **Data access and processing requires network authentication**
- **Server Hosted Database Solutions - use registered TCP/IP port**
  - Users authenticated at server-level, not client-level
  - Encrypted network traffic - Server uses TripleDES encryption with HMACSHA-I for integrity checking (168 bit key)
  - Additional 128 bit SSL between FileMaker server and client
- **VPN is required if accessing from outside of network**
- **e-Data entry verified via record-level audit log**
- **Unique usernames with passwords required**
- **Logon/logoff whenever switching users**

Questions?

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER

34