Generating, Accessing and Processing of Biomedical Data

Davit Sargsyan
Associated Director, Statistics

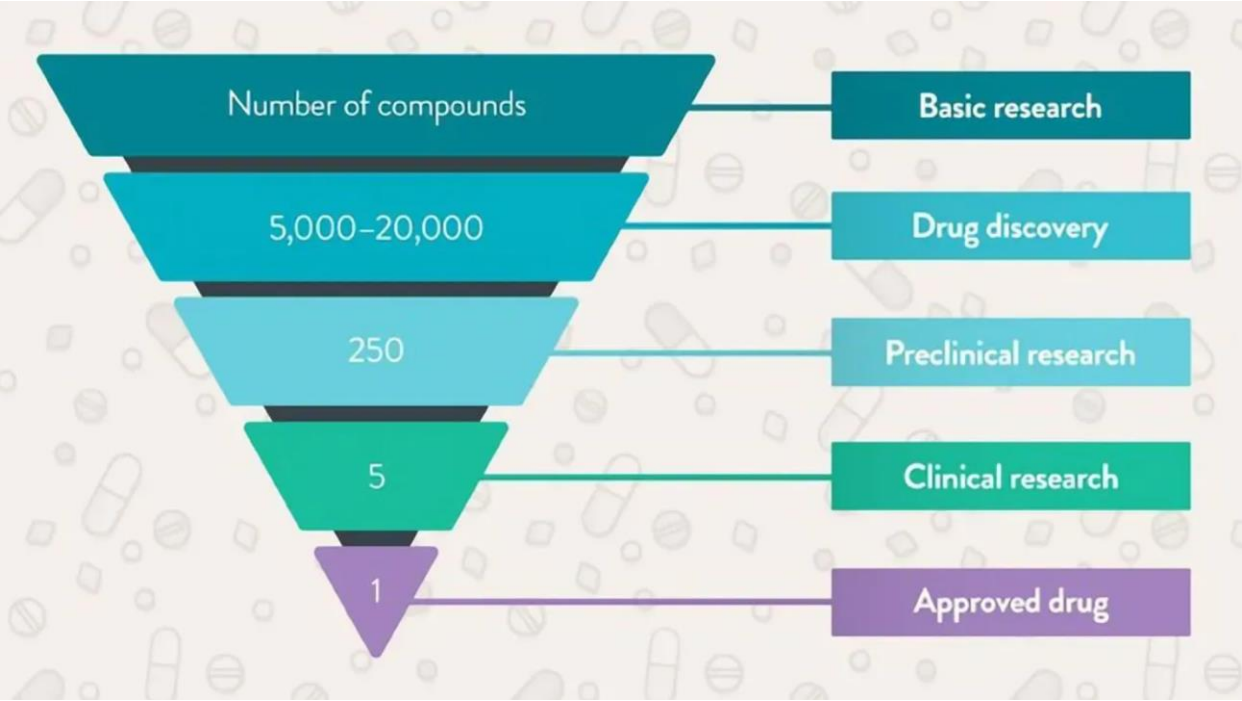
June 7th, 2022

Pavetis

Table of Contents

Drug Development Overview

Drug Development Phases

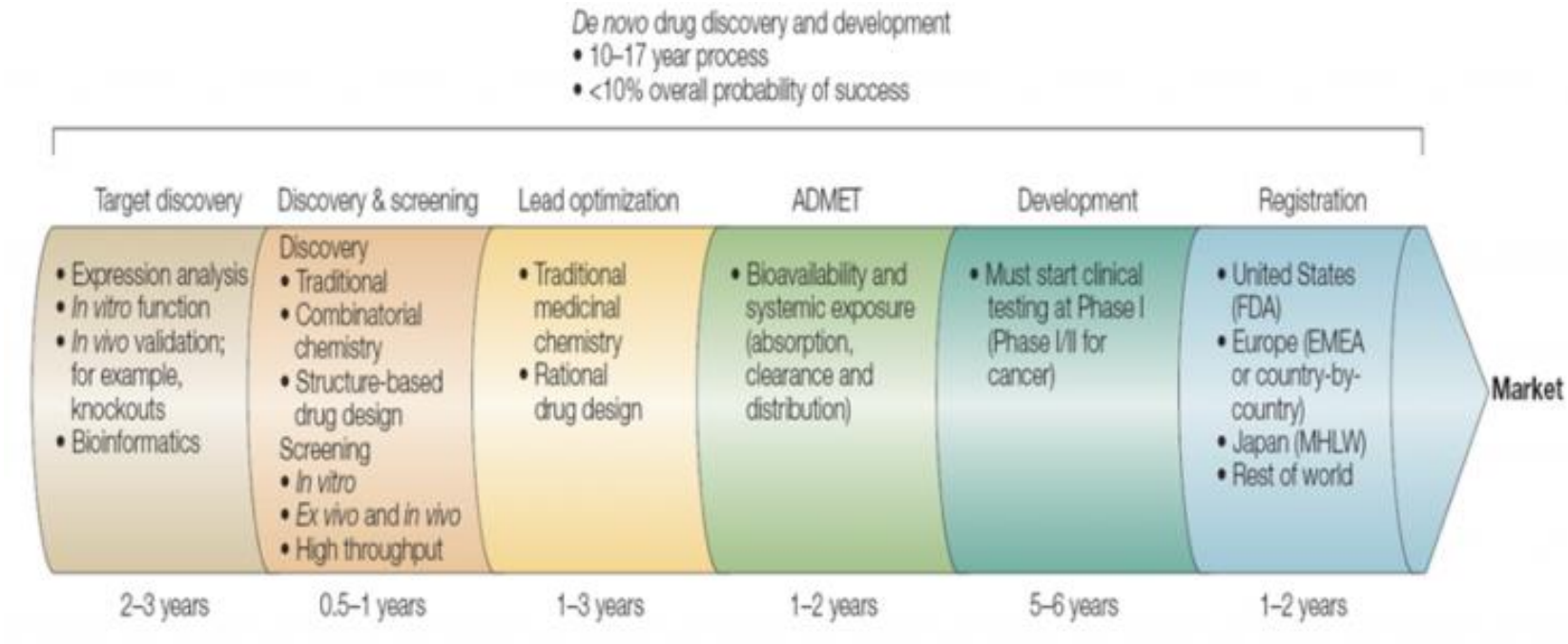


- In the US, academia is working closely with the industries.
- The most laborious part of the research is typically conducted in universities
- Once successful product is identified, a commercial company can be spun off. The university helps with patents and resources but retains partial ownership.
- If the new company produces promising products, it can be acquired by a larger company or raises capital to bring the product to market

https://www.technologynetworks.com/drug-discovery/articles/exploring-the-drug-development-process-331894

Drug Development Timeline

- It takes about $2B to develop a new drug
- The research does not stop with the market lunch. Post-market monitoring, continuous safety and efficacy studies in the population, and possible new indications for the drug continue for decades



https://www.ebi.ac.uk/training/online/courses/functional-genomics-i-introduction-and-design/drug-discovery/

Nonclinical Studies

Compound Discovery and Testing

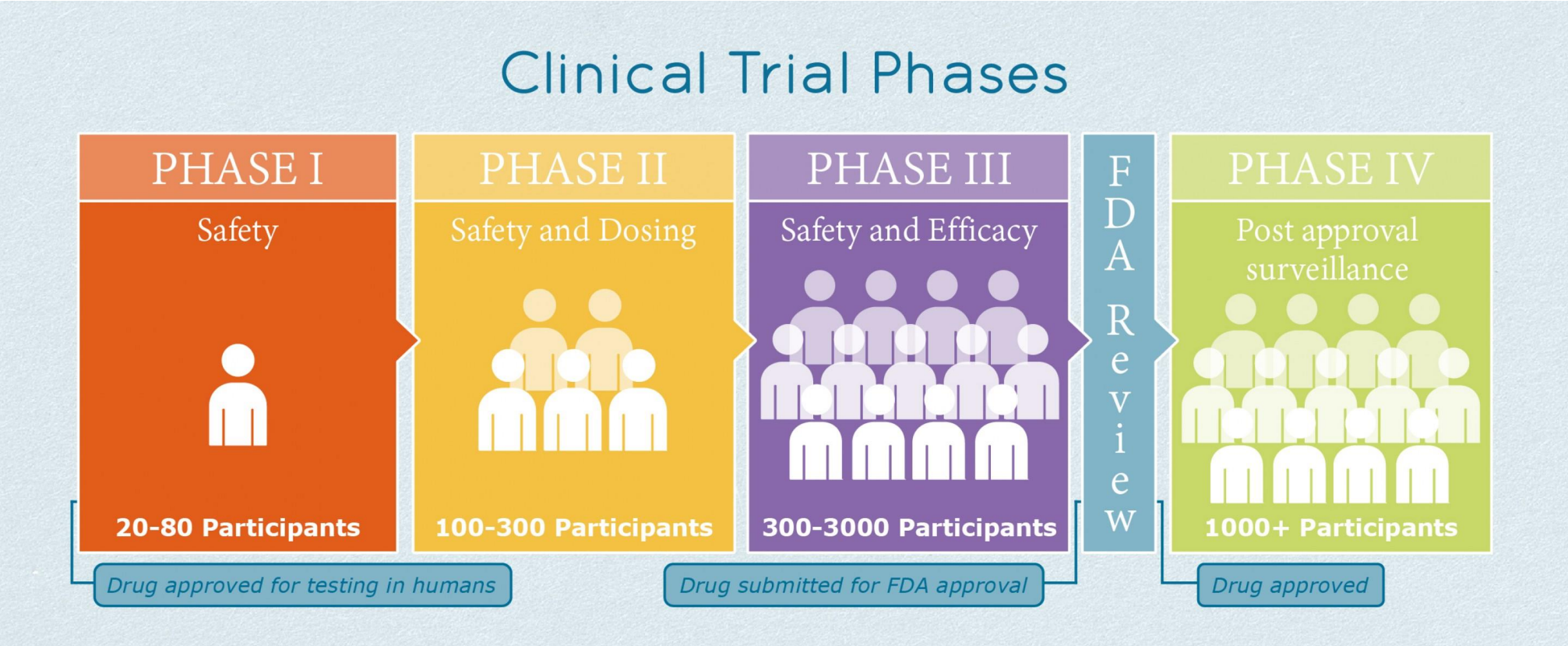
- Natural extracts from plans, proteins from bacteria or animals, chemically synthesized molecules, etc. are isolated and tested
- In vitro (cells in petri dishes), in vivo (animals) and post-vivo experiments test and confirm biological activity of the compound

Types of Studies

- Exploratory: large number of compounds are screened
- Confirmatory: the effects must be consistent so several studies should show confirm efficacy and non-toxicity

Formulation and Manufacturing

- Determine correct dose (pharmacodynamic)
- Determine best delivery route (injection, oral, etc.)
- Packaging and other inactive ingredients affect pharmacokinetics (half-life, clearance, etc.)
- Quality control of manufacturing process
- Drug stability (determines expiration date)



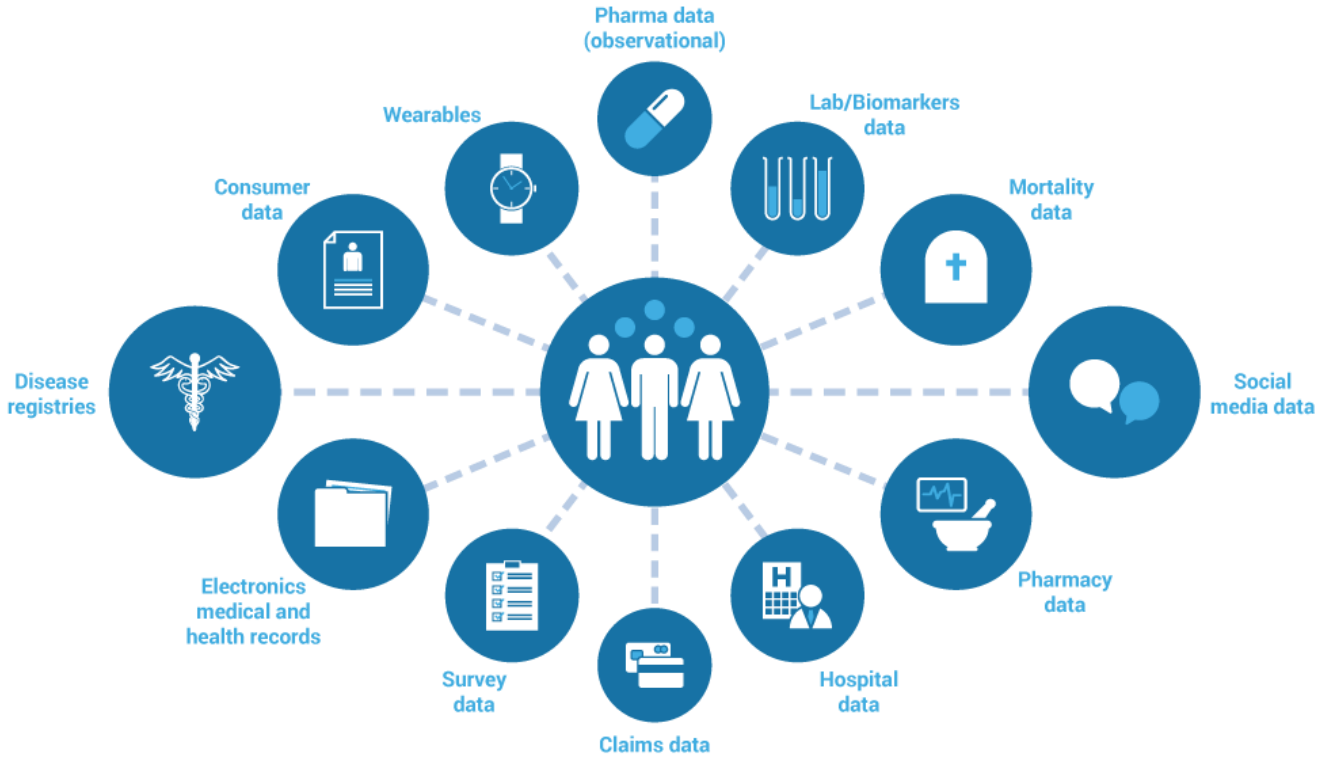
https://www.ildcollaborative.org/resources/phase-iv-ipf-clinical-trials

- In the US, before a new drug goes to market, it must be approved by the Food and Drugs Administration (FDA)
- The drug can the be marketed and sold (but only for the approved indication!)

Real World Evidence

- The US does not have a unified database for patients. Instead, data is collected by many independent institutions:

- Hospitals collect admission and discharge data
- Insurance companies collect claims data including doctor visits, procedures, lab work and medication information
- States can aggregate data from multiple hospitals and healthcare systems (groups of hospitals, insurance companies, etc.)
- National Institute of Health (NIH) and other federal health organizations can create their own data bases: patient level or aggregates



https://www.chcuk.co.uk/china-key-considerations-in-using-real-world-evidence-to-support-drug-development/

- Research organizations request, buy, process and curate many of these databases

Design of Experiment, Data Collection and Formatting

Experimental Design Considerations

- What are the objectives of my study?
- What is the criteria for success?
- What is the primary hypothesis I am testing?
- What is the minimal change in the primary endpoint that I will consider to be biologically significant? Is it achievable with the current design?
- How much variability should I expect? What are the main sources of variability in this type of experiments? Search literature, use in-house data and previous experience.
- What are the appropriate statistical models?

Research Question & Hypotheses

Available Resources

- Animals
- Budget
- Compounds
- Supplies

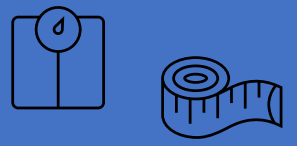
Variables

- Dependent/Outcome
- Independent/Predictor

Appropriate Statistical Test/Model

How Variables Will Be Measured

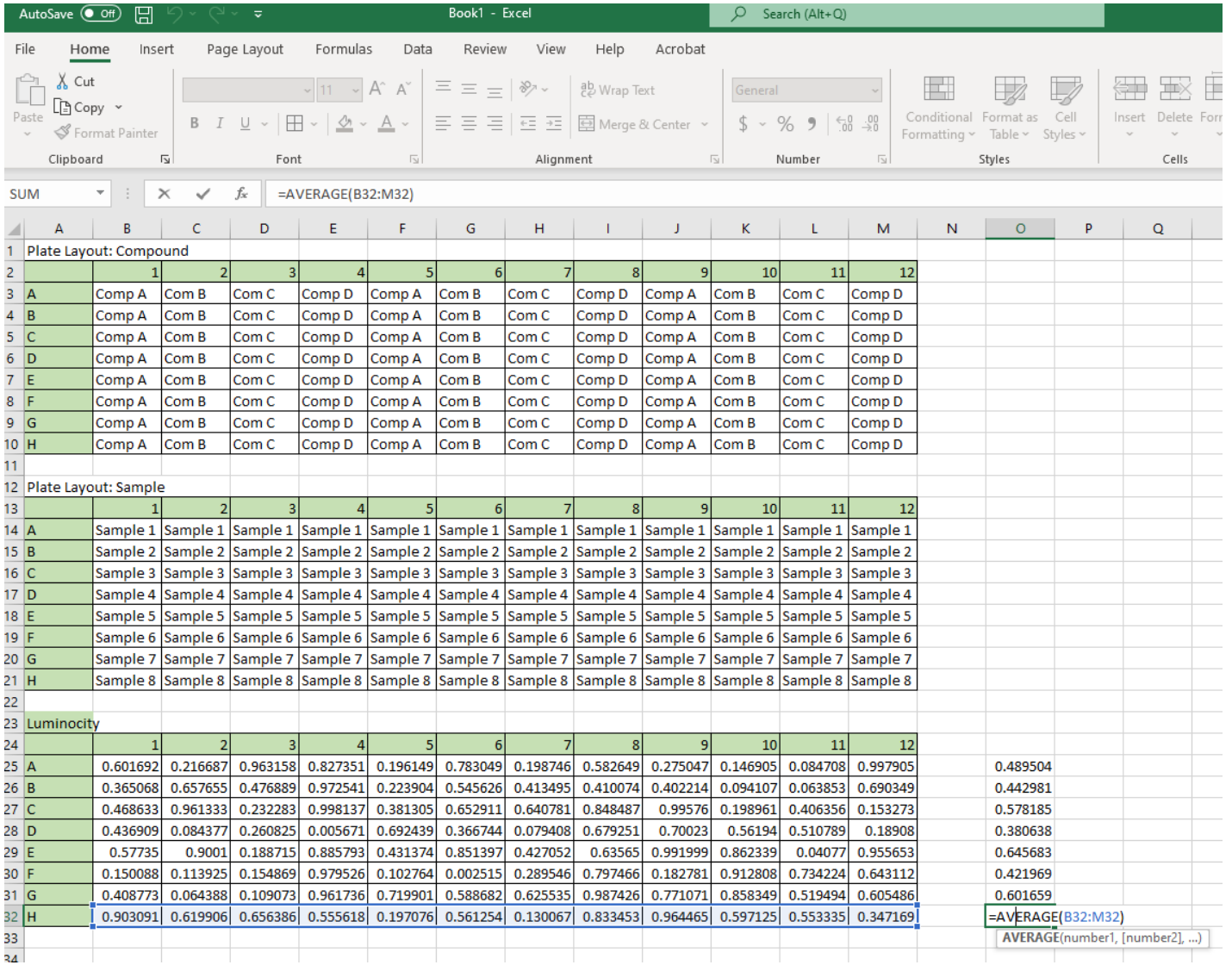
- Metric
- Units



Inference

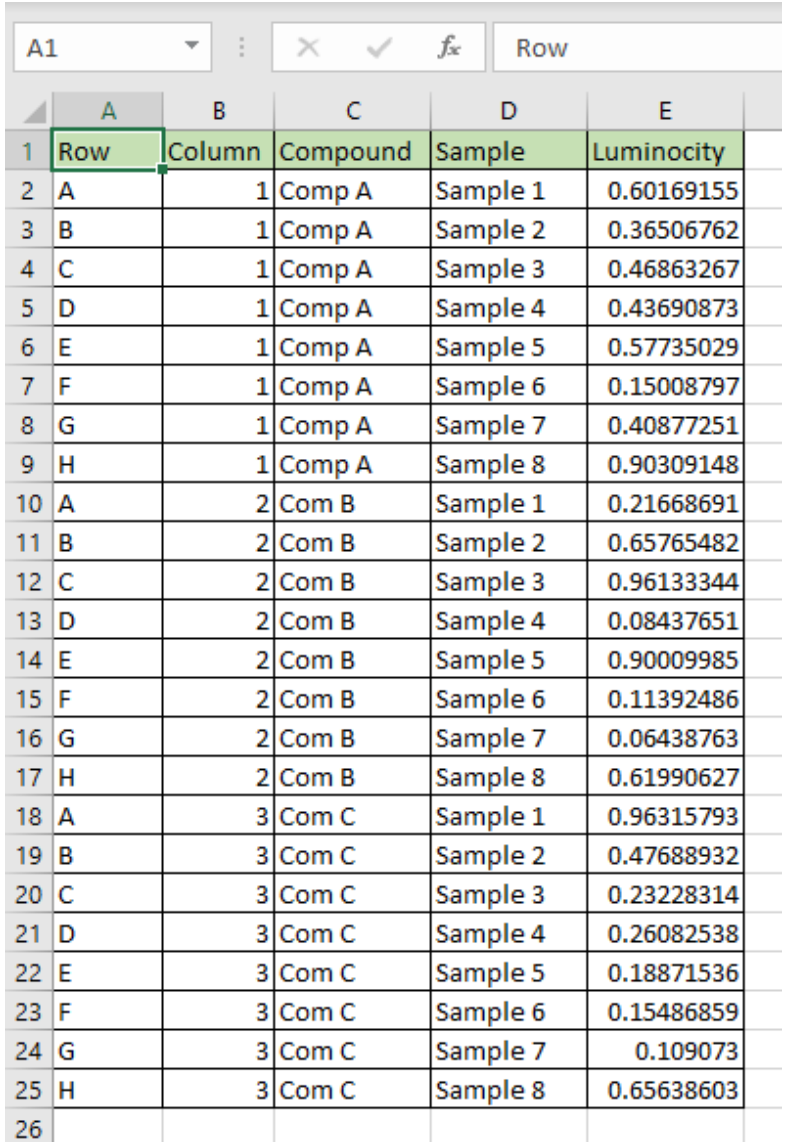
Data Collection in MS Excel

- Preclinical experiments are typically small (less than 100 animals) but the amount of data collected can be huge (genomics, proteomics, microbiome, biomarkers, etc.)
- Microsoft Excel is often used to record basic information about the experiment: treatment groups, experimental conditions, genotype
- Many scientific instruments can output data in Excel or text format (.txt, .csv, .tsv)
- Excel also allows basic data processing, analysis and visualization
- Be careful with massive Excel Workbooks containing a lot of formulas!



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Plate Layout: Compound																
2		1	2	3	4	5	6	7	8	9	10	11	12				
3	A	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
4	B	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
5	C	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
6	D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
7	E	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
8	F	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
9	G	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
10	H	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D	Comp A	Com B	Com C	Comp D				
11																	
12	Plate Layout: Sample																
13		1	2	3	4	5	6	7	8	9	10	11	12				
14	A	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1	Sample 1				
15	B	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2	Sample 2				
16	C	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3	Sample 3				
17	D	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4	Sample 4				
18	E	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5	Sample 5				
19	F	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6	Sample 6				
20	G	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7	Sample 7				
21	H	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8	Sample 8				
22																	
23	Luminocity																
24		1	2	3	4	5	6	7	8	9	10	11	12				
25	A	0.601692	0.216687	0.963158	0.827351	0.196149	0.783049	0.198746	0.582649	0.275047	0.146905	0.084708	0.997905		0.489504		
26	B	0.365068	0.657655	0.476889	0.972541	0.223904	0.545626	0.413495	0.410074	0.402214	0.094107	0.063853	0.690349		0.442981		
27	C	0.468633	0.961333	0.232283	0.998137	0.381305	0.652911	0.640781	0.848487	0.99576	0.198961	0.406356	0.153273		0.578185		
28	D	0.436909	0.084377	0.260825	0.005671	0.692439	0.366744	0.079408	0.679251	0.70023	0.56194	0.510789	0.18908		0.380638		
29	E	0.57735	0.9001	0.188715	0.885793	0.431374	0.851397	0.427052	0.63565	0.991999	0.862339	0.04077	0.955653		0.645683		
30	F	0.150088	0.113925	0.154869	0.979526	0.102764	0.002515	0.289546	0.797466	0.182781	0.912808	0.734224	0.643112		0.421969		
31	G	0.408773	0.064388	0.109073	0.961736	0.719901	0.588682	0.625535	0.987426	0.771071	0.858349	0.519494	0.605486		0.601659		
32	H	0.903091	0.619906	0.656386	0.555618	0.197076	0.561254	0.130067	0.833453	0.964465	0.597125	0.553335	0.347169		=AVERAGE(B32:M32)		
33															AVERAGE(number1, [number2], ...)		
34																	

Data Formatting



A1 | Row

	A	B	C	D	E
1	Row	Column	Compound	Sample	Luminocity
2	A	1	Comp A	Sample 1	0.60169155
3	B	1	Comp A	Sample 2	0.36506762
4	C	1	Comp A	Sample 3	0.46863267
5	D	1	Comp A	Sample 4	0.43690873
6	E	1	Comp A	Sample 5	0.57735029
7	F	1	Comp A	Sample 6	0.15008797
8	G	1	Comp A	Sample 7	0.40877251
9	H	1	Comp A	Sample 8	0.90309148
10	A	2	Com B	Sample 1	0.21668691
11	B	2	Com B	Sample 2	0.65765482
12	C	2	Com B	Sample 3	0.96133344
13	D	2	Com B	Sample 4	0.08437651
14	E	2	Com B	Sample 5	0.90009985
15	F	2	Com B	Sample 6	0.11392486
16	G	2	Com B	Sample 7	0.06438763
17	H	2	Com B	Sample 8	0.61990627
18	A	3	Com C	Sample 1	0.96315793
19	B	3	Com C	Sample 2	0.47688932
20	C	3	Com C	Sample 3	0.23228314
21	D	3	Com C	Sample 4	0.26082538
22	E	3	Com C	Sample 5	0.18871536
23	F	3	Com C	Sample 6	0.15486859
24	G	3	Com C	Sample 7	0.109073
25	H	3	Com C	Sample 8	0.65638603
26					

- Multiple tables/tables in Excel are useful for quickly going through the data but are not easy to read into analytical tools
- A better way to store analysis datasets is to melt it into a single long table. Data from each original table can be represented by a data column
- If data is produced by an instrument or collected into a template, write an adapter code that reshapes the original data into an analysis dataset format

NOTE: from here on I will mainly refer to R programming language and RStudio environment for data analysis but same goes for SAS, SPSS, GraphPad Prism, Python and so on

SEND Initiative (Nonclinical Data)

- Preclinical and small clinical studies often follow their own convention for data collection and storage
- This “Wild Wild West” of data formats is prevalent in academia and in the industry alike
- Standard for Exchange of Nonclinical Data (SEND) initiative started in 2002. Since 2016 the FDA accepts raw toxicology data in this format.
- SEND is a standard data model to store various types of preclinical data. Table and variable names are standardized.
- Unfortunately, unlike clinical data models (discussed later), SEND is not widely adopted

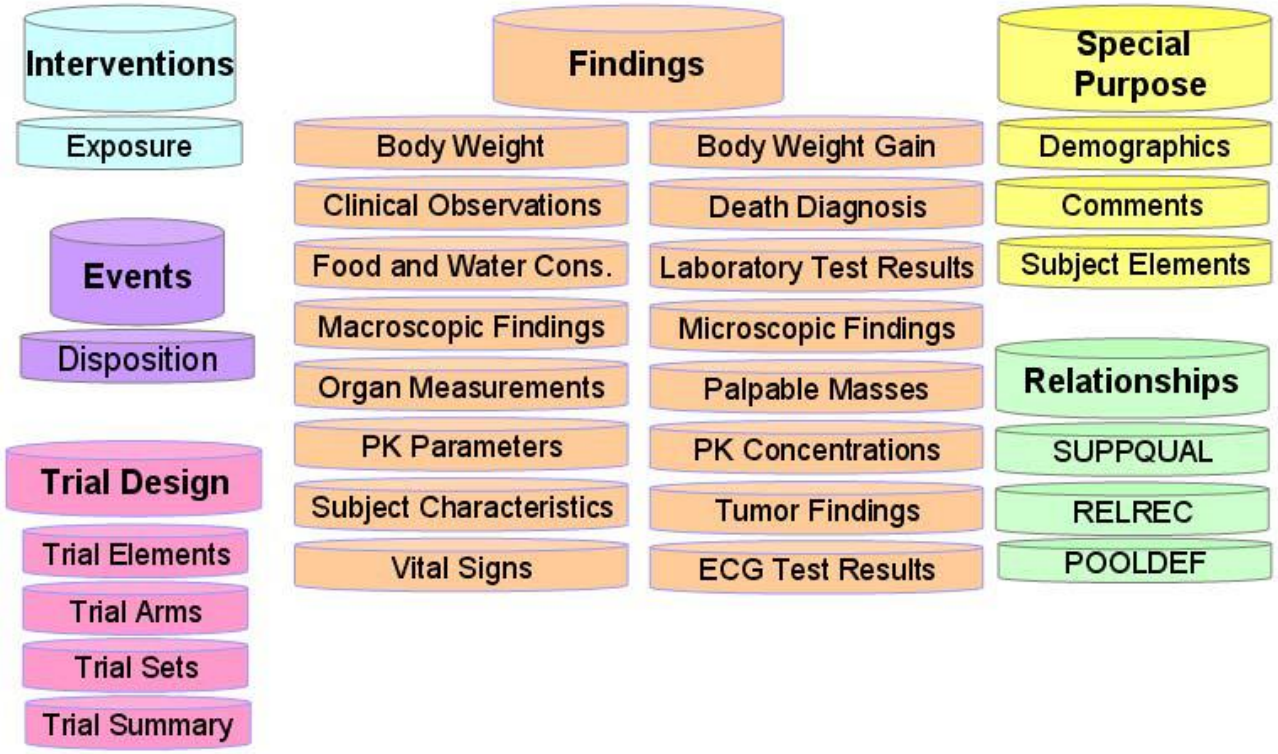


Figure 1: SENDING observational classes and special-purpose domains. *The Standard for the Exchange of Nonclinical Data (SEND): History and Basics*. Wood, F; Kramer, L A. PharmaSUG 2011 Paper CD14

SDTM (Clinical Data)

- Study Data Tabulation Model (SDTM) – standard structure for clinical (human) trials and foundation for SEND
- Defined by the Submission Data Standards team of Clinical Data Interchange Standards Consortium (CDISC)
- Data divided into Domains (standard datasets), grouped into 3 classes: Interventions, Events, or Findings
- Domains are abbreviated with 2 letters (CO, DM, …). Variable names are no longer than 8 letters. Lookup in data dictionary.
- Many other data models: Analysis Data Model (ADaM), Laboratory Data Model (LAB), Digital Imaging and Communications in Medicine (DICOM)

Special-Purpose Domains:
•Comments (CO)
•Demographics (DM)
•Subject Elements (SE)
•Subject Visits (SV)

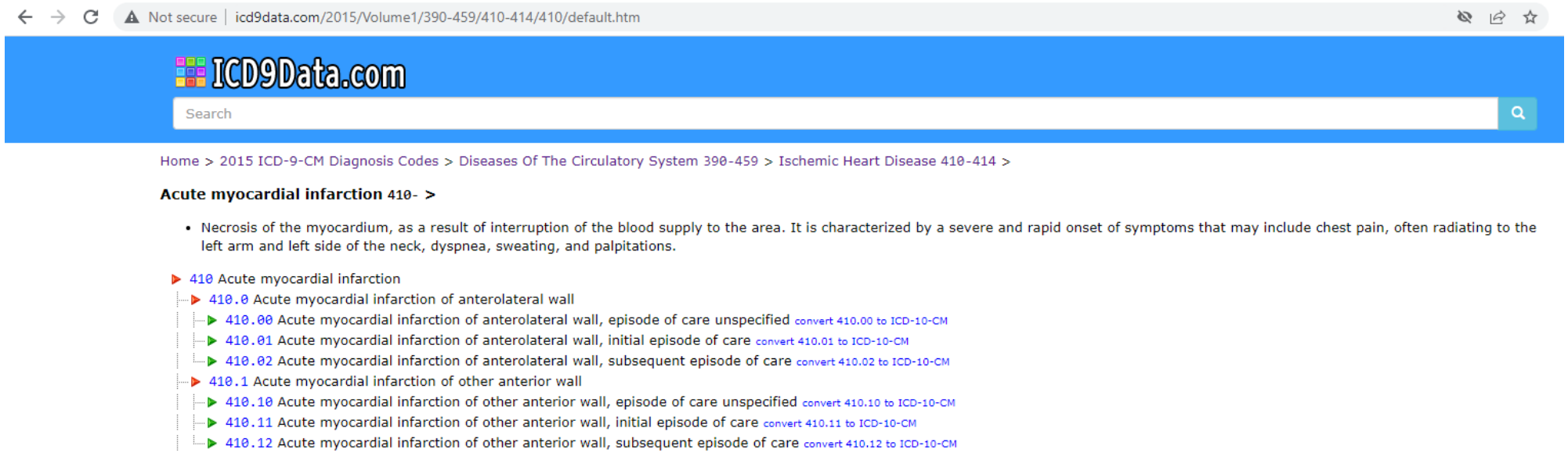
Interventions General Observation Class:
•Concomitant Medications (CM)
•Exposure as Collected (EC)
•Exposure (EX)
•Substance Use (SU)
•Procedures (PR)

Events General Observation Class:
•Adverse Events (AE)
•Clinical Events (CE)
•Disposition (DS)
•Protocol Deviations (DV)
•Medical History (MH)
•Healthcare Encounters (HO)

Findings General Observation Class:
•Drug Accountability (DA)
•Death Details (DD)
•…

Medical Coding: ICD

- Many observational datasets have much simpler structure (e.g., a flat file containing a single table) but use medical coding systems
- International Classification of Diseases (ICD) codes are maintained by World Health Organization (WHO) and used for epidemiology, health management and billing (hospitals, insurance companies, etc.)
- Current version is ICD-11 but the US only switched to ICD-10 in 2015
- Hierarchical model: ~13,000 billable codes in ICD-9 and ~68,000 in ICD-10
- Easy code search: http://www.icd9data.com/



Coded Dataset Example

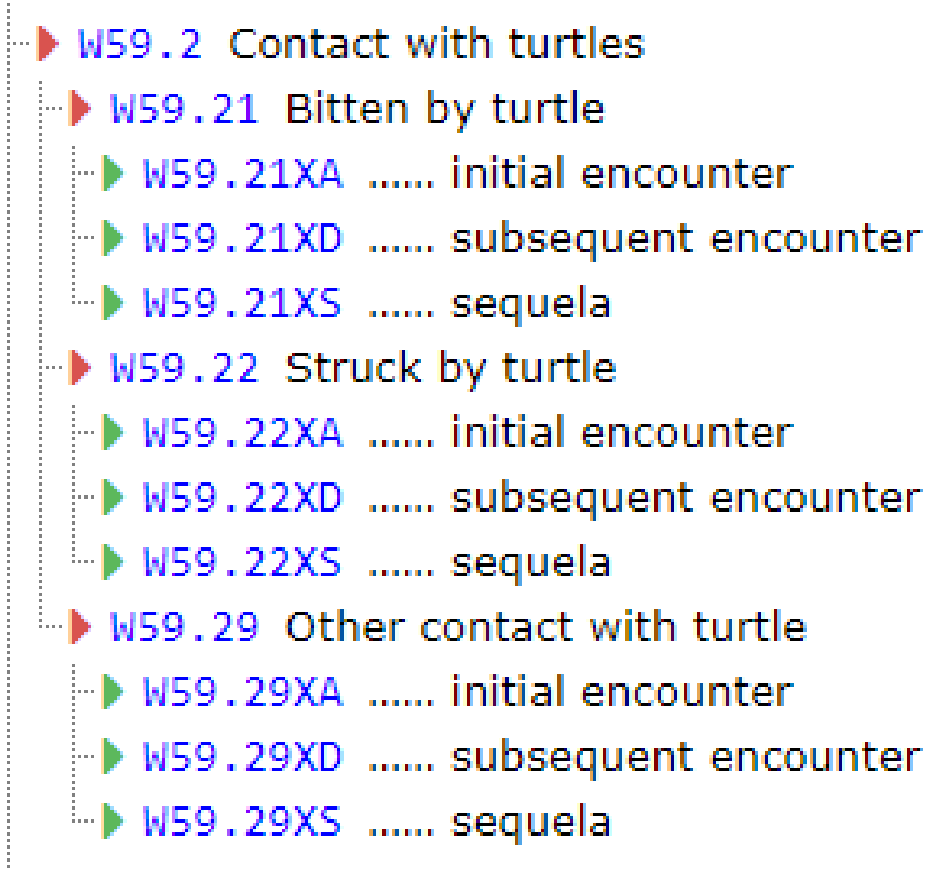
ICD-9 Dataset Example

- Example of ICD-9 coded dataset is the Myocardial Infarction Data Acquisition system (MIDAS)

- Contains ~17M records of ~4M patients admitted to NJ hospitals between 1995 and 2015 (data after 2015 is ICD-10-coded and has not been fully merged yet)

Aside: Sometimes “less is more”. ICD-10 system is much more detailed which also makes it harder to use. Some ICD-10 codes were made fun of.

See more of funny ICD-10 codes here and here.



Synthetic Data Based on MIDAS

	Patient_ID	patbdte	NEWDTD	SEX	PRIME	DX1	DX2	DX3	DX4	DX5	DX6	DX7	DX8	DX9	ADMDAT
1	1	1966-01-14	2013-10-01	F	Commercial	7218	65233	78440	19882	11595	9632	1228	E8338	9331	2003-05-31
2	1	1966-01-14	2013-10-01	F	Commercial	01010	8941	71888	01402	2590	6279	82521	63502	E8409	2008-12-15
3	2	1965-07-26	NA	F	Commercial	3510	1467	71894	6200	E9389	01384	NA	NA	NA	2005-06-26
4	2	1965-07-26	NA	F	Commercial	29515	36216	36363	NA	NA	NA	NA	NA	NA	2008-04-16
5	2	1965-07-26	NA	F	Commercial	80235	71874	0401	NA	NA	NA	NA	NA	NA	2013-11-22
6	2	1965-07-26	NA	F	Commercial	7500	37501	NA	NA	NA	NA	NA	NA	NA	2015-01-12
7	3	1932-07-19	2015-10-26	M	Medicare	37863	20692	1228	65810	V8301	86359	0075	E8132	E8749	1994-11-25
8	3	1932-07-19	2015-10-26	M	Medicare	67333	E8429	E0141	6220	67512	8441	8291	NA	NA	1997-01-12
9	3	1932-07-19	2015-10-26	M	Medicare	67333	5836	6387	73301	34672	73303	NA	NA	NA	2001-12-24
10	3	1932-07-19	2015-10-26	M	Medicare	36459	96901	66550	37181	4412	3151	64272	NA	NA	2003-04-09
11	3	1932-07-19	2015-10-26	M	Medicare	71885	5185	2331	64763	71697	V171	3638	65421	52131	2006-05-10
12	3	1932-07-19	2015-10-26	M	Medicare	7337	01184	66391	1505	65561	6387	NA	NA	NA	2007-04-11
13	3	1932-07-19	2015-10-26	M	Medicare	4761	1505	5679	82000	NA	NA	NA	NA	NA	2010-05-22
14	3	1932-07-19	2015-10-26	M	Medicare	1725	43831	94514	5723	NA	NA	NA	NA	NA	2011-05-16
15	3	1932-07-19	2015-10-26	M	Medicare	52434	25092	0771	80372	66702	NA	NA	NA	NA	2013-02-15
16	4	1951-07-08	NA	F	Medicaid/Self-Pay/Other	E8508	V8289	NA	NA	NA	NA	NA	NA	NA	2010-12-14
17	4	1951-07-08	NA	F	Medicaid/Self-Pay/Other	8712	51635	3643	85196	NA	NA	NA	NA	NA	2011-07-07
18	5	1935-03-01	NA	F	Medicare	2821	94514	86810	NA	NA	NA	NA	NA	NA	2005-02-27

Converting ICD-9 Codes to Variables

Filter

	Patient_ID	patbdte	ADMDAT	NEWDTD	SEX	PRIME	uc	ami	stroke	tia	chf	hyp	diab	asc	copd	lipid	cld	akd	ckd	cmyo
						column 4: unknown														
1	1	1941-03-01	1999-01-21	NA			1	1	0	0	0	0	0	0	0	0	0	0	1	0
2	2	1968-04-25	1997-07-11	NA	F	Commercial	1	1	0	0	0	1	0	0	0	0	1	1	1	0
3	3	1925-05-04	2004-08-09	2012-10-22	M	Medicaid/Self-Pay/Other	1	1	0	0	1	0	1	0	0	1	0	0	0	0
4	4	1944-01-22	2009-11-22	NA	M	Commercial	1	1	0	0	0	1	0	0	0	1	0	0	0	0
5	5	1935-12-04	2004-01-02	NA	F	Medicare	1	1	0	0	0	0	0	1	0	0	0	0	0	0
6	6	1940-06-07	2010-04-22	NA	F	Medicare	1	1	0	0	0	0	0	0	0	0	0	0	0	0
7	7	1923-02-28	1995-04-26	2003-12-10	F	Commercial	1	0	1	0	1	0	1	1	0	0	1	0	1	1
8	8	1946-10-20	2015-04-12	NA	F	Medicare	1	0	0	1	0	1	0	0	0	0	0	0	0	0
9	9	1926-10-12	2005-04-21	NA	F	Commercial	1	0	0	1	0	0	0	1	0	0	1	0	0	0
10	10	1969-06-23	2011-01-28	2012-09-13	F	Commercial	1	1	0	0	1	0	1	0	0	1	0	0	0	1
11	11	1964-09-14	1997-05-14	NA	M	Medicare	1	1	0	0	0	0	0	0	0	0	0	0	1	0
12	12	1948-04-25	1997-11-12	NA	F	Commercial	1	1	0	0	0	0	0	1	0	0	0	0	0	1
13	13	1947-05-12	2010-03-24	2015-11-17	M	Commercial	1	1	0	0	0	0	0	1	0	0	0	0	0	0
14	14	1955-05-31	2003-12-26	NA	M	Medicare	1	1	0	0	1	0	1	0	1	0	1	0	0	0
15	15	1927-10-03	2009-01-14	NA	M	Medicare	1	1	0	0	0	1	0	0	0	0	0	1	0	0
16	16	1967-11-02	1995-07-17	2006-09-06	F	Medicare	1	0	1	0	0	1	0	0	0	1	0	1	1	0
17	17	1934-05-25	1998-11-29	2002-09-14	M	Medicare	1	0	1	0	0	0	0	0	0	0	0	0	0	0

- Each variable (outcome, comorbidity) must be defined using one or more ICD-9 codes
- Often, timing is considered, e.g., an outcome that occurred after the main event (HF after AMI), or preexisting condition (Diabetes before HF)
- Statisticians and clinicians work together to define each condition

Practice: Data Import and Processing In R

Thank You

Contact
Davit Sargsyan
Cardiovascular Institute of NJ, Rutgers University
sargdavid@gmail.com

Navetis