Statistical Collaboration

Department of Neurology Faculty Meeting

Fei Ye, PhD, MSPH | Associate Professor

Co-Director of Collaborative Studies Coordinating Center

Department of Biostatistics, Vanderbilt University Medical Center

Vanderbilt Center for Quantitative Sciences

615-936-5169 | fei.ye@vanderbilt.edu

Collaboration Plan

- Study design (experimental design, power analysis/sample size, interim analysis, randomization schedule, etc)
- Statistical/bioinformatic analysis (data harmonization, descriptive analysis, data visualization, statistical testing, model development and validation, etc)
- Statistical consulting
- Manuscript preparation and grant application (addressing statistical /bioinformatics issues, responding to reviewers' comments, writing relevant paragraphs)

Collaboration Plan

- Contact us: fei.ye@Vanderbilt.edu
- Senior staff statistician:
  - Run Fan, MS in Biostatistics; PhD in Microbiology

- Statisticians are in high demand! ☺
  - We help multiple faculty members in the department as well as VUMC collaborators outside the department, so please always give advance notice (at least 2 months for grant applications).
  - For long-term projects and complex analyses, it makes sense to allocate %efforts to biostatisticians (me and Run, or another pair of faculty and staff biostatisticians) from your funded study so we could devote sufficient amount of time to your project to ensure the quality of our work.

How Can You Help Us?

An example data dictionary (not perfect)

Variable Name	Coding	Comments
Birthday	yyyymmdd	Subject's birth date
Age	Year, decimal	Age at sample collection *DERIVED VARIABLE*
Sex	0 = Female 1 = Male	Subject's biological sex
Education	1 = Elementary school or less 2 = More than elementary school but less than high school 3 = High school 4 = More than high school 10 = Missing	Highest level of education completed *DERIVED VARIABLE*
Height	cm	Measured at physical exam
Weight	kg	Measured at physical exam
BMI	kg/m^2	Subject's body mass index *DERIVED VARIABLE*
BMI_WHO	1 = Underweight (<18.5 kg/m^2) 2 = Normal (18.5 - 24.9 kg/m^2) 3 = Overweight (25.0 - 29.9 kg/m^2) 4 = Obese (≥30.0 kg/m^2)	BMI categorization, according to the WHO universal scale *DERIVED VARIABLE*

One Way to Make Our Life Miserable…

First name	Last name	D.O.B (dd/mmm/yy)	Sex	Centre	Unique I.D Number	sex	Date of Diagnosis	Primary Tumor Site	Site (head and neck, trunk, extremities,	Stage at diagnosis (after complete	Histologic subtype	Breslow Depth (mm)	Mitoses / HPF
abc	*xyz*	*31453*	*M*	*MIA*	*MIA 1*								
cde	*fgh*	*31818*	*M*	*VDB*	*VDB1*								
jkl	*pqr*	*25639*	*F*	*MDA*	*MDA1*								
	25086091		M	*VDB*	V1	male	38706	cutaneous (neck)	head and neck	IIIc (pTxpN3pMx)	Nodular	n/a	
	29223021		M	*VDB*	V2	female	39812	cutaneous (back)	trunk			4 to 5	
	30596068		M	*VDB*	V3	Male	39767	unknown	unknown	Stage IV			
	35943158		M	*VDB*	V58	Female	39696	Cutaneous (left up	trunk	2b		no data	
	33190364		M	*VDB*	V59	Female	39457	Cutaneous (scalp	head and neck		nodular/ssm	2.65 (nodular); 0.45 (ssm)	
	32602914		M	*VDB*	V60	male	40575	unknown	unknown				
	36709319		M	*VDB*	V61	F	40835	Cutaneous (upper	trunk	3	nodular	7.5	4
	36898310		F	*VDB*	V95	F	41194	Cutaneous (right f	extremity	Ia		0.4	<1
	12357448		F	*VDB*	V96		33317	Cutaneous (left up	trunk	unknown		unclear	
	22572317		M	*VDB*	V97	M	41333	Cutaneous (anore	mucosal	IIIb		2.5	2

		D.O.B (dd/mmm/yy) If not able to provide, enter age at timepoints column J,AE	Sex	Centre	Unique I.D Number	Mutation Status (include major relevant genes tested e.g. BRAF/NRAS	Date of Diagnosis of Stage IV (dd/mmm/yy) Leave blank if only IIIC	Comment		Drug 1 name	Best Response (CR, PR, SD, PD)	Drug 2 name	Best Response (CR, PR, SD, PD)	Drug 3 Name	Best Response (CR, PR, SD, PD)	Comment
		24-Jul-50	M	MDACC	MDACC1	NRAS Q61L/BRAF WT/KIT WT	5-Jun-2014			Ipilimumab	PD	x		x		
		17-Oct-58	F	MDACC	MDACC2	NRAS Q12S/BRAF WT/KIT WT	9-Apr-2014			Cisplatin/Velban/Temozolomide	PD	Ipilimumab	PD	x		
		21-Mar-62	M	MDACC	MDACC3	NRAS G12E/BRAF WT/KIT WT	##########			HDIL-2	x	x		x		

How Can You Help Us?

- Create your data dictionary
  - Before collecting data, write a detailed list of the information to be collected and the concepts to be measured in the study.
  - For each variable, the data dictionary should include at least: a unique variable name, variable label, type of variable (numeric, categorical, logical, etc), permissible values and/or range of values, additional edits to be performed for logic checks.

Please Treat Your Statistician Nicely by:

- Using standard data structure where
  - Each row corresponds to an individual subject (or unit of analysis)
  - Each column corresponds to a different variable or measurement.
  - One row per subject.
  - Use same formatted values for same levels of a variable, check for upper/lower-case and typos (e.g., female, male, Female, MALE, 1, 0, femail …)
  - Use same note for missing values (99, missing, NA) or just leave it blank
- Using the first row of the spreadsheet for unique variable names (good idea to avoid very long names and special symbols)
- When sending new data, do NOT change variable names (i.e., column names)

- Assign each subject (or unit of analysis) a unique ID (eg, 1, 2, 3, etc).
- No text (other than NA for missing values) should be entered in a column intended for numbers. Don’t mix text and numbers in the same column

When multiple data files are generated from a study:

- Every record in every data file must contain a unique subject (or unit of analysis) ID that is consistent across all files.
- Data files that are likely to be merged should not use the same variable names (other than the common ID variable).

- For data with repeated measurements on the same subject,

  Two options: a “wide” data file or a “long” data file. For 20 patients, each contributes 5 data points (measured at 5 different time points):

  - “Wide”: 20 rows and 6 columns (5 blood pressures and an ID). Still have one record (row) per subject.

  - “Long”: 100 rows of 3 columns (ID, week number (1-5), and blood pressure). Have 5 records (rows) per subject.

  - Long form is often preferred but which option to use will depend on the target statistical program.

Other Resources

- VICTR
- Summer Institute
- Collaborative Studies Coordinating Center https://www.vumc.org/cscc/