
Biostatistics and Study Design

Patrick Ten Eyck, PhD

Assistant Director for Biostatistics and Research Design

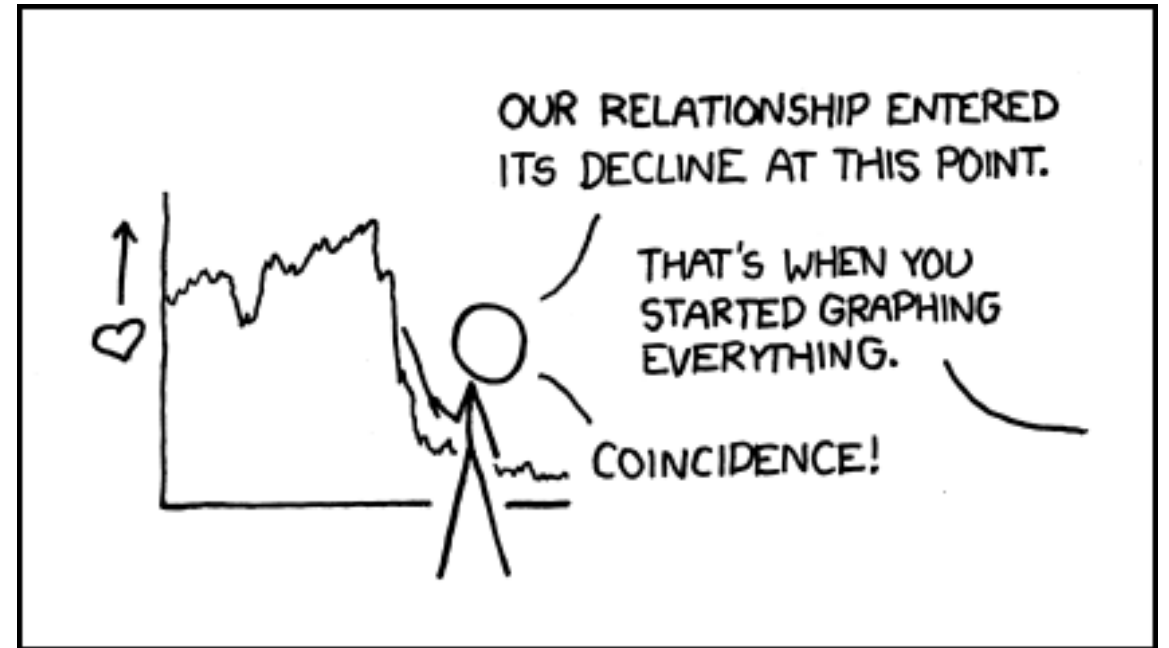
December 18, 2018

Disclosure

Within the past 12 months, I have had no financial relationships with proprietary entities that produce health care goods and services.

The Steps in a Study

1. Plan the study
2. Conduct the experiment
3. Clean the data
4. Run initial analyses
5. Perform planned analyses



Step 1 – Plan the Study

- Formulate a hypothesis
 - Will the answer to the question be relevant to medicine, to explain a biological principle or mechanism, or to establish drug efficacy or safety?
 - Can the hypothesis be proven, given the constraints involved in patient-oriented research?

Step 1 – Plan the Study

- Consider the study population
 - Want to include patients who have relevant medical condition but not those who might be “too sick” for the study (i.e., focus on mild manifestations of disease)
 - Are there certain inclusion/exclusion criteria that must be considered? (e.g., age, sex, medications, other conditions, etc.)
 - Can a representative sample be enrolled and recorded?
 - What are the costs associated with recruiting and retaining subjects?
 - What is our expected proportion of dropout subjects?

Step 1 – Plan the Study

- Contact a statistician IMMEDIATELY!
 - They can assist you throughout the study so earlier is better
 - If you are in the early stages of your study, we can run power and sample size calculations, provide data collection strategies, and generate an analysis plan
 - Inaccuracies that occur will carry through to the end of the study



Step 1 – Plan the Study

- Things to keep in mind for the statistics consultation:
 - Prepare a brief overview of the study.
 - Have an idea of what your data collection sheet will look like. This may impact interpretation at the end of the analysis.
 - Research if similar studies have been conducted before. If so, is there a publication you could provide so that we can learn more about the statistical methods used and if they should be used in your study?

Step 1 – Plan the Study

- Determine the nature of the data. For each variable, is it a:
 - Response variable?
 - Identify the outcome(s) of interest.
 - How many are there?
 - What type of variable are they?
 - Continuous or categorical?

Step 1 – Plan the Study

- Determine the nature of the data. For each variable, is it a:
 - Predictor variable?
 - What is the main predictor that is being tested? (group, treatment, etc.)
 - Are there other variables (controls, confounders, etc.) we should consider in the model?
 - We don't want to fit models that are missing crucial information. This leads to inaccurate results!

Step 1 – Plan the Study

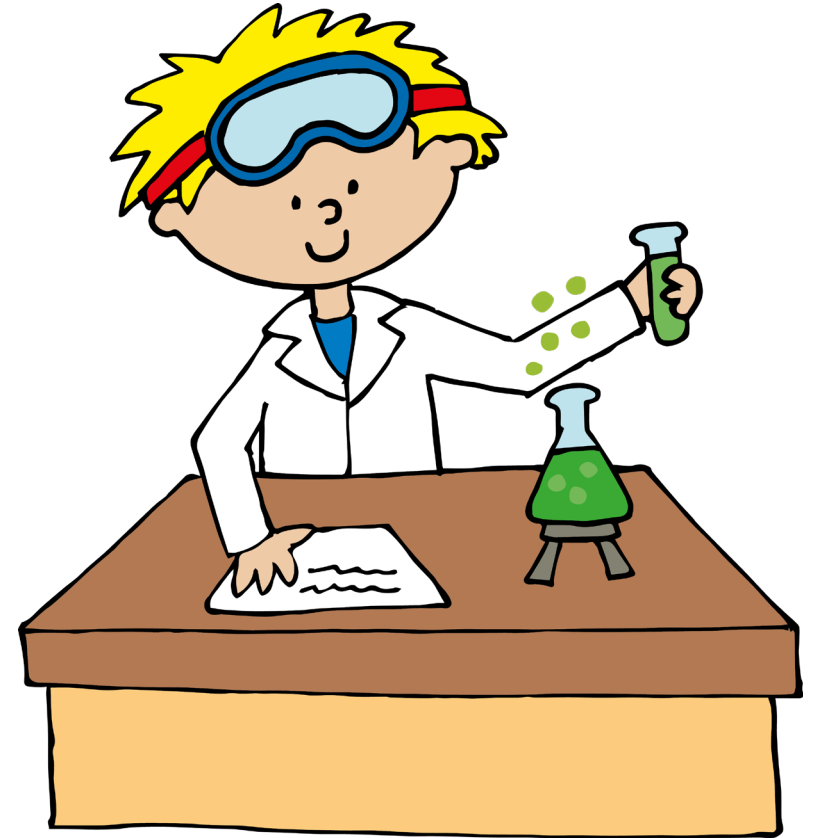
- Determine the nature of the data. For each variable, is it a:
 - Response variable?
 - Identify the outcome(s) of interest.
 - How many are there?
 - What type of variable are they?
 - Continuous or categorical?

Step 1 – Plan the Study

- Determine what kind of results you are looking for.
 - Effect size estimates, confidence intervals, hypothesis tests, plots, etc.
 - Constructing shell tables will help with this since you will be able to see what measures you can model and how they can relate.
 - Statisticians can always provide suggestions on how best to disseminate study findings to the rest of the scientific community but you know what you want to show.

Step 2 - Conduct the Experiment

- Enroll the subjects
- Take measurements
- Enter into data capture program (REDCap, Access, Excel, etc.)



Step 3 - Clean the Data

- This is the lion's share of a statistician's work so try to provide data that requires as little cleaning as possible.
- Frequently, data files contain formatting that makes sense to the investigator, but is unreadable by statistical software programs.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Patient serial	Zyflo Start dat	Zyflo end date	Time from star	Test Date		FEF25-75% POST	FEF25-75% PRE			FEV1/FVC POST	FEV1/FVC PRE		FVC POST	FVC PRE
2	1	2/5/2007	death		6/10/2010		0.57	0.85			62	65		2.82	2.51
3					3/21/2010		0.70	0.78			62	65		2.88	2.55
4					12/5/2006		0.68	0.85			59	65		2.63	2.58
5					8/6/2007	6	0.57	0.38			63	59		2.85	2.44
6					8/15/2008		0.73	0.43			62	59		2.46	2.14
7					3/6/2009		0.66	0.80			62	64		2.86	2.64
8					1/22/2010		0.77	0.80			62	64		2.78	2.57
9					9/3/2010		0.63	0.62			60	62		2.78	2.57
10	2	1/29/2007	0		1/29/2007		0.47	0.44			52	57		1.30	1.36
11					2/1/2007		0.96	0.51			56	45		1.64	1.12
12					1/2/2008		1.69	1.13			62	54		5.22	5.68
13	4	6/6/2007	4/21/2009		6/25/2008	11	0.59	0.63			63	64		2.00	1.90
14					6/6/2007		0.73	0.76			68	69		1.88	1.86
15					8/29/2007	2	0.76	0.59			69	66		1.78	1.49
16					1/28/2008	5	0.73	0.74			69	68		1.82	1.84
17					2/13/2009		0.65	0.45			67	63		1.58	1.45
18	6	10/1/2008	12/1/2010		2/15/2008		1.29	1.88			66	70		3.88	3.75
19					11/6/2009		4.48	4.61			82	81		5.37	5.52
20					3/11/2009		4.61	5.61			83	86		5.25	5.08
21					1/15/2010	5	3.54	4.08			79	81		4.91	4.83
22					9/1/2010		2.89	2.75			73	74		5.24	5.16
23					3/2/2011		2.73	2.24			73	71		5.03	4.88
24	7	5/20/2011	12/1/2010		12/9/2011		0.76	0.82			60	60		2.67	2.59
25					5/3/2013		0.74	0.70			63	59		2.25	2.27
26					5/20/2011		0.73	0.88			64	59		3.08	3.16
27					6/22/2011		0.81	0.76			62	60		2.37	2.45
28					9/9/2011		0.95	0.70			64	58		2.62	2.59
29					4/20/2012	4	0.81	0.78			63	60		2.43	2.37
30					8/10/2012	15	0.89	0.87			64	63		2.36	2.47
31					12/7/2012		0.81	0.80			61	59		2.39	2.52
32					9/27/2013		0.78	0.74			61	59		2.61	2.58
33	8	3/24/2010	12/1/2010		10/31/2012		0.66	0.75			55	58		3.50	3.27
34					6/12/2013		0.76	0.82			61	59		2.89	2.70
35					11/8/2013		0.76	0.65			57	53		3.78	3.88
36					3/24/2010		0.65	0.43			54	46		3.37	3.28
37					5/5/2010		0.62	0.65			54	55		3.37	3.03
38					8/18/2010	5	0.73	0.63			58	57		3.85	3.85
39					12/1/2010	9	0.77	0.64			60	57		3.39	3.06

Step 3 - Clean the Data

- Remove extraneous notes, rows, columns, cells with color filled in, etc. It is best to make a copy of the original, uncleaned sheet so you can always reference back to it.
- The more carefully the study plan was generated, the less time-consuming this process is.

	A	B	C	D	E	F	G	H	I	J	K
1	Patient se	Zyflo Start date	Zyflo end date	Trt Group	Test Date	FEF25-75%	FEF25-75%	FEV1/FVC	FEV1/FVC	FVC POST	FVC PRE
2	1	2/5/2007	death	WCG	6/10/2010	0.57	0.85	62	65	2.75	2.41
3	1	2/6/2007	death	WCG	3/21/2010	0.7	0.78	62	65	2.82	2.51
4	1	2/7/2007	death	WCG	12/5/2006	0.58	0.45	59	62	2.86	2.55
5	1	2/8/2007	death	WCG	8/6/2007	0.57	0.38	59	55	2.63	2.58
6	1	2/9/2007	death	WCG	8/15/2008	0.73	0.53	63	59	2.65	2.41
7	1	2/10/2007	death	WCG	3/6/2009	0.66	0.43	62	59	2.46	2.14
8	1	2/11/2007	death	WCG	1/22/2010	0.77	0.8	62	64	2.86	2.64
9	1	2/12/2007	death	WCG	9/3/2010	0.63	0.62	60	62	2.78	2.57
10	2	1/29/2007		DDQ	1/29/2007	0.47	0.44	32	37	4.3	3.36
11	2	1/30/2007		DDQ	2/1/2007	0.96	0.51	56	45	4.64	4.12
12	2	1/31/2007		DDQ	1/2/2008	1.69	1.13	62	54	5.22	4.58
13	4	6/6/2007	4/21/2009	DDQ	6/25/2008	0.59	0.63	63	64	2	1.9
14	4	6/6/2007	4/21/2009	DDQ	6/6/2007	0.73	0.76	68	69	1.88	1.86
15	4	6/6/2007	4/21/2009	DDQ	8/29/2007	0.76	0.59	69	66	1.78	1.49
16	4	6/6/2007	4/21/2009	DDQ	1/28/2008	0.73	0.74	69	68	1.82	1.84
17	4	6/6/2007	4/21/2009	DDQ	2/13/2009	0.65	0.45	67	63	1.58	1.45
18	6	10/1/2008	12/1/2010	WCG+DDQ	2/15/2008	1.29	1.68	66	70	3.88	3.75
19	6	10/1/2008	12/1/2010	WCG+DDQ	11/6/2009	4.48	4.61	82	81	5.37	5.52
20	6	10/1/2008	12/1/2010	WCG+DDQ	3/11/2009	4.61	6.01	83	86	5.25	5.08
21	6	10/1/2008	12/1/2010	WCG+DDQ	1/15/2010	3.54	4.08	79	81	4.91	4.83
22	6	10/1/2008	12/1/2010	WCG+DDQ	9/1/2010	2.89	2.75	73	74	5.24	5.16
23	6	10/1/2008	12/1/2010	WCG+DDQ	3/2/2011	2.73	2.24	73	71	5.03	4.88
24	7	5/20/2011	12/1/2010	WCG	12/9/2011	0.76	0.82	60	60	2.57	2.59
25	7	5/20/2011	12/1/2010	WCG	5/3/2013	0.74	0.7	63	59	2.25	2.27
26	7	5/20/2011	12/1/2010	WCG	5/20/2011	0.73	0.59	64	59	2.08	2.15
27	7	5/20/2011	12/1/2010	WCG	6/22/2011	0.81	0.76	62	60	2.37	2.45
28	7	5/20/2011	12/1/2010	WCG	9/9/2011	0.95	0.7	64	58	2.52	2.59
29	7	5/20/2011	12/1/2010	WCG	4/20/2012	0.81	0.78	63	60	2.43	2.37
30	7	5/20/2011	12/1/2010	WCG	8/10/2012	0.89	0.87	64	63	2.36	2.47
31	7	5/20/2011	12/1/2010	WCG	12/7/2012	0.81	0.8	61	59	2.39	2.52
32	7	5/20/2011	12/1/2010	WCG	9/27/2013	0.78	0.74	61	59	2.51	2.58
33	8	3/24/2010	12/1/2010	WCG+DDQ	10/31/2012	0.66	0.75	55	58	3.5	3.27
34	8	3/24/2010	12/1/2010	WCG+DDQ	6/12/2013	0.76	0.82	61	59	2.89	2.7
35	8	3/24/2010	12/1/2010	WCG+DDQ	11/8/2013	0.76	0.65	57	53	3.78	3.88
36	8	3/24/2010	12/1/2010	WCG+DDQ	3/24/2010	0.55	0.43	54	45	3.37	3.28
37	8	3/24/2010	12/1/2010	WCG+DDQ	5/5/2010	0.62	0.65	54	55	3.37	3.03
38	8	3/24/2010	12/1/2010	WCG+DDQ	8/18/2010	0.73	0.63	58	57	3.85	3.85
39	8	3/24/2010	12/1/2010	WCG+DDQ	12/1/2010	0.77	0.64	60	57	3.39	3.06

Step 4 - Run Initial Analyses

- Use distributions and scatterplots to visualize data.
- You will have a better understanding of each individual variable, as well as how particular variables relate in a basic sense. This is important information that is needed for the data analysis.
- This is when you calculate your Table 1 descriptive statistics.
 - You can also provide p-values, which assess the significance of various predictor variables between the main treatment groups.
 - Any variables that are unbalanced in distribution between treatment groups may need to be analyzed further in a multivariate setting.

Step 5 - Perform Planned Analyses

- Use initial analyses to check assumptions for statistical inference (hypothesis testing, point/interval estimation).
- Calculate model information criteria if model selection is involved.
- Fill in the shell tables.
- Interpret the results.



Missing Data

- Missing completely at random (MCAR)
 - The probability of missing observation is not dependent on any observed and/or any missing value
- Missing at random (MAR)
 - The probability of missingness is dependent only on the observed outcome and completely independent of any unobserved values
- Missing not at random (MNAR)
 - The probability of missingness is dependent also on the missing values instead of just the observed outcomes.

Handling Missing Data

- Complete case analysis
 - Ignore observation from analysis
- Last observation carry forward (LOCF)
 - Missing values replaced by value from previous observation
- Mean imputation
 - Unrealistic assumption, can yield biased estimates even under MCAR
 - Can exaggerate treatment effects and inflate type I error
- Multiple imputation
 - Involves combining estimates and standard errors obtained from multiple copies of the original data set in which missing values are replaced by randomly generated values based on a predictive model

Multiple Comparisons

- If we test multiple parameters (or a single parameter over multiple conditions), we need to adjust our decision criteria so that we stay below the type I error rate.
- One of the simpler approaches is the Bonferroni correction:
 - For k parameters being tested, each test will be compared to $\alpha^* = \frac{\alpha}{k}$
 - This is the simplest correction but it somewhat conservative

Analysis Time!

- Now that we are all experts on the steps involved in a study, let's run through an analysis.
- This analysis will compare the burnout rates of two different groups of Pediatrics physicians (fellows & junior faculty).
- Several other variables will be used in this analysis, including: age, gender, marital status, regular exercise, weekly hours on duty, and personal debt.
- So what should our next step be?

Contact a statistician!!!



How Should You Introduce Your Study?

- General goal: Do current fellows at UI Peds Dept have higher Maslach Burnout Inventory scores compared to faculty within 3 years of graduation in same Dept?
- Specific goal: Show that a difference in burnout (continuous or dichotomized MBI) exists between these groups. Specifically, the pediatric fellows at UI have higher burnout scores/rates compared to junior faculty.
- Outcome variable: Burnout is quantified continuously by the Maslach Burnout Inventory (MBI), but can also be made dichotomous by specifying a cutoff value for the MBI, usually 27. We will need to check the distribution of MBI to dictate our exact analytical approach.
- Predictor variable(s): The main predictor will be current position but we may want to control for age, gender, marital status, exercise, overnight calls, and financial debt.

Analysis Plan

- First, we will calculate the sample size needed. We will use $\alpha = 0.05$, power = 0.8, and assume equal sample sizes. Based on existing information on burnout, our belief is that fellows & junior faculty have:

- Mean MBIs of 28.5 and 26.4, respectively, each with a standard deviation of 2.5.

N total = 60 (or 30 per group)

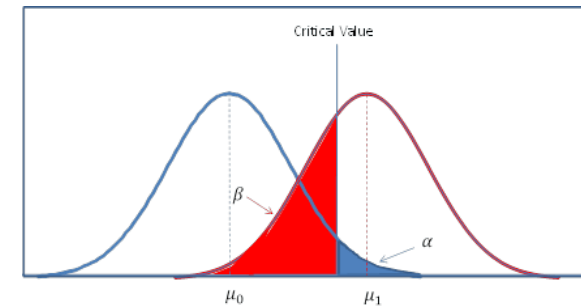
- Burnout proportions of 50% and 40%, respectively.

N total = 666 (or 333 per group)

- For the analysis, we will model two outcomes:

- Continuous, normally distributed MBI
- Dichotomous MBI (1: MBI ≥ 27 , 0: MBI < 27)

- We want to do an initial assessment of the variables in the data set so we should generate a shell table of descriptive statistics and measures of significance for differences between groups.



Shell Table for Descriptive Statistics

	Fellow (N=30)	Faculty (N=30)	P-value (parametric)	P-value (non-parametric)
Age (years)				
Female (%)				
Married (%)				
Exercise (hours/week)				
Overnight calls (hours/week)				
Debt (\$)				
Burnout score				
Burnout \geq 27				

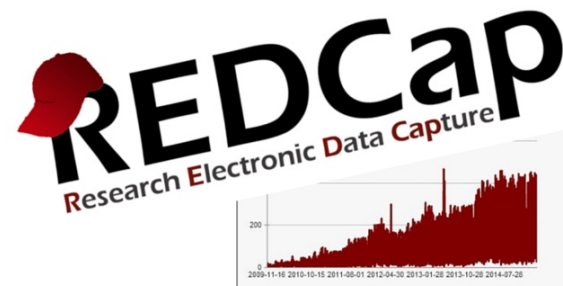
Additional Modeling

- For both outcome variables, we will fit univariate models of the predictors. This will provide us with an unadjusted comparison.
- Then, for both outcome variables, we will fit full models (i.e., all predictors simultaneously).
- We can then compare the unadjusted and adjusted modeling estimates and p-values.

	Burnout Normally Distributed		Burnout Dichotomous	
Variable	Univariate	Multivariate (full model)	Univariate	Multivariate (full model)
Group				
Age				
Female				
Married				
Exercise				
Overnight calls				
Debt				

Conduct the Study

- Now that you've confirmed the analysis plan, it is time to conduct the study.
- Enroll your subjects.
- Gather the data.
- Enter the data into the capture program.



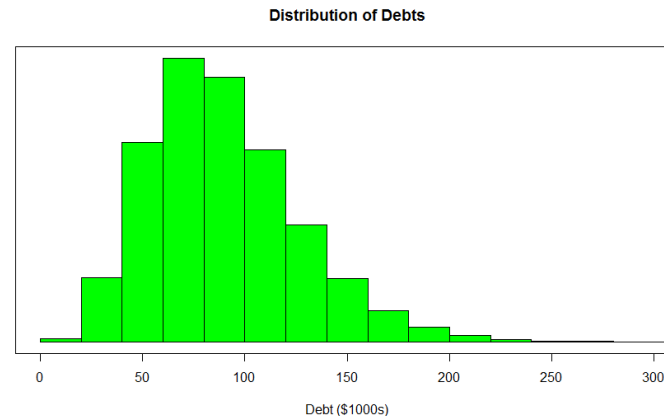
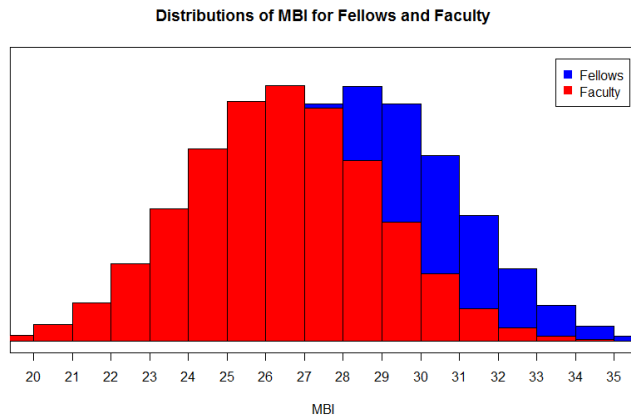
Clean the Data

- Make sure you provide the data in a format that the statistical software can read.
- This process varies in time to complete and can require frequent communication to make sure it has been done accurately.
- REDCap will help make this easier.



Run Initial Analyses

- Make plots of important variables and fill out the demographic table. This will help us identify variables that could be significant to the multivariate models.



Frequency Percent Row Pct Col Pct	Table of Group by Married			
	Group	Married		Total
		Yes	No	
Fellow	20 33.33 66.67 60.61	10 16.67 33.33 37.04	30 50.00	
Faculty	13 21.67 43.33 39.39	17 28.33 56.67 62.96	30 50.00	
Total	33 55.00	27 45.00	60 100.00	

Shell Table for Descriptive Statistics

	Fellow (N=30)	Faculty (N=30)	P-value (parametric)	P-value (non-parametric)
Age (years)	30.2 (1.6)	36.3 (2.2)	0.004	0.026
Female (%)	57.5%	60.0%	0.846	0.915
Married (%)	30.6%	55.9%	0.078	0.127
Exercise (hours/week)	7.2 (1.5)	4.0 (1.2)	0.071	0.114
Overnight calls (hours/week)	4.6 (0.8)	6.2 (2.3)	0.210	0.345
Debt (\$)	92,500 (15,000)	48,500 (20,000)	< 0.001	< 0.001
Burnout score	28.5 (2.7)	26.4 (2.4)	0.002	0.037
Burnout \geq 27	51.9%	38.4%	0.198	0.257

Perform Planned Analyses

	Burnout Normally Distributed	
Variable	Univariate	Multivariate (full model)
Fellow vs. Faculty	2.1 (0.002)	1.8 (0.017)
Age	0.214 (0.067)	0.017 (0.267)
Female	-0.176 (0.867)	0.642 (0.301)
Married	2.30 (0.023)	1.76 (0.078)
Exercise	-0.891 (0.036)	-0.578 (0.067)
Overnight calls	0.613 (0.147)	0.678 (0.126)
Debt (\$1000s)	0.017 (0.152)	0.008 (0.264)

Mean difference
(p-value)

Continuous Conclusions

- Multivariate linear modeling on the normally distributed outcome variable MBI reveals several significant factors from our variables under consideration. It is expected that fellows will have a higher MBI than junior faculty by an expected 1.8 units ($p = 0.017$). Those who are married expect a higher MBI by 1.76 units ($p = 0.078$) compared to their unmarried counterparts. Each additional hour of exercise per week leads to a reduction in MBI by 0.578 units ($p = 0.067$). All estimates are adjusted for the full model set of covariates.

Perform Planned Analyses

	Burnout Dichotomous	
Variable	Univariate	Multivariate (full model)
Fellow vs. Faculty	1.523 (0.126)	1.316 (0.203)
Age	1.002 (0.267)	1.001 (0.421)
Female	0.984 (0.912)	1.079 (0.457)
Married	1.143 (0.068)	1.092 (0.197)
Exercise	0.957 (0.256)	0.973 (0.561)
Overnight calls	1.037 (0.305)	1.041 (0.284)
Debt (\$1000s)	1.001 (0.244)	1.001 (0.267)

Odds Ratio
(p-value)

Dichotomous Conclusions

- Multivariate linear modeling on the dichotomous outcome variable $MBI \geq 27$ reveals no statistically significant factors from our variables under consideration. The only variable that was marginally significant was married status (only for the univariate setting). As determined in the study plan, the sample size used in this portion of the analysis provides insufficient power. In order to more accurately determine the significance of these factors on high MBI status, we will have to rerun this study with a larger sample size.

Biostatistics Services at Iowa



The Biostatistics Core Alliance provides a common point of access to biostatistical support for members of the Holden Comprehensive Cancer Center (HCCC), the Institute for Clinical & Translation Science (ICTS), and other biomedical programs at the University of Iowa.

<https://www.public-health.uiowa.edu/bca/>

Biostatistics Services at Iowa

- Alliance members are highly skilled biostatisticians from the:
 - College of Public Health (BCC & CPHS)
 - HCCC
 - ICTS
- These centers provide expertise in:
 - Data management
 - Study design
 - Statistical analysis of clinical, epidemiological, laboratory, and public health data
 - Grant and manuscript preparation.
- A request for biostatistical support can be made via the accompanying online form and will be routed to a Core member with expertise relevant to the request.

Biostatistics Services through the ICTS

• Pre-analysis

- Calculate sample size
- Conduct power analysis
- Provide data collection strategies
- Generate analysis plan
- Manage data



• Analysis

- Conduct data analysis
- Consolidate results into interpretable form



• Post-analysis

- Develop methods/results sections for paper or presentation



Getting Started

- All collaborations begin with a FREE one-hour consultation
- This allows us to:
 - Discuss the aims of the study and form an initial analysis plan
 - Determine if additional services are needed (services can be added later by requestor or consultant)
 - Assess complexity of analysis and appropriate staff and rates
 - Establish an approximate timeline based on urgency

Additional Services Include:



POWER ANALYSIS/ SAMPLE SIZE CALCULATION

Determine statistical power or sample size required for research study



RESEARCH DESIGN AND STUDY PREPARATION

Provide data collection/analysis strategies to best answer research questions of interest



DATA MANAGEMENT/ ANALYSIS

Manage data sets involved in study, conduct analysis, condense results into interpretable form



REPORTING

Develop analysis section for grants, papers, and presentations

Service Rates



SENIOR BIOSTATISTICIAN

\$75/hour



JUNIOR BIOSTATISTICIAN

\$50/hour

Requesting Biostats Services



I-CART

<https://i-cart.icts.uiowa.edu>



CLICK

on Institute for
Clinical and
Translational
Science, then
Biostatistics



SELECT


Select “Add” next to
FREE One-Hour
Biostatistical
Analysis
Consultation

Thank you!

Questions?

Institute for Clinical and Translational Science

Patrick Ten Eyck

 SW44-M GH
200 Hawkins Drive
Iowa City, IA 52242

 319-384-5242

 icts.uiowa.edu

 facebook.com/ICTSIowa

 [@ICTSIowa](https://twitter.com/ICTSIowa)

 patrick-teneyck@uiowa.edu