Omics Data Clinic Notes and Analyses

2025 April 22

Nicole Kirchoff, Immunity, Infection, and Inflammation

We are interested in how clinical variables are associated with antibiotic resistance genes from stool samples of NICU babies. We have repeated measures as stool samples have been taken from babies at multiple timepoints.
I believe a generalized linear mixed model would be most appropriate to analyze our data. However, I would like some help with glmm interpretation. How should I decide on the best background distribution (negative bionomial, poisson, etc.)? Also, how can we be confident that the fixed effects are important in shaping the abundance of antibiotic resistance genes? Notably, the R package lme4 does not have p-values in some of its glmm results (I understand p-values can be controversial here). Should I not be paying attention to p-values, but other ways of interpreting these results?

2025 April 15

Jeeyeon Cha, Medicine/Endocrinology

Advice on statistical power calculations for 1) mutant mouse experiments and 2) cell culture experiments. For Mouse experiments, we will isolate pancreatic islets from WT and mutant groups, males and females, and perform multiomics analysis. For cell culture experiments, we will obtain primary human pancreatic islets and genetically mutate them to have multiple groups and perform multiomics analysis.

2024 November 19

Romario Lobban (Leon Bellan), Mechanical Engineering

We are working on cooling-triggered drug delivery. Recently, we have collected in vivo data showing increased release of a dye (meant to model a drug) upon cooling the animal. To quantify dye release, we analyzed fluorescence signal. We would like help determining the statistical significance of the difference in fluorescence signal between the cooled and uncooled animals.

2024 November 12

Kimberly Bress (Carissa Cascio), Psychiatry and Behavioral Sciences

This project is investigating the relationship between facial expressivity data and functional connectivity metrics from rs-fMRI.
I want to test the hypothesis that individuals within our sample can be clustered into distinct groups based on principle components extracted from timeseries data, specifically changes in facial expressivity metrics sampled rapidly over a 3 minute task. I think want to evaluate whether there are differences in functional connectivity with a specific brain network across these clusters.
Hypothesis: Individuals can be clustered into distinct groups based on their action unit probability scores over time, and this clustering is associated with differences in functional connectivity in the facial sensorimotor network.
Analysis plan:

1. First, run a PCA with the timeseries data, the probability scores for each AU at each timepoint in the MET task – the inputs for the PCA will include the timeseries for each AU (the change in the AU likelihood score over the course of the entire task). There are 20 total action units.
2. Run a clustering analysis using the components from the PCA
3. Construct a mixed effects model with the following variables:
   i) DV = Z-values from FC matrix
   ii) IV1 = cluster label
   iii) IV2 = group
   iv) IV3 = gender
   v) IV4 = age
   vi) IV5 = group * cluster label
   vii) RE1 = subject

2024 September 10

Angela Liu (Shawniqua Williams Roberson), Internal Medicine

Studying EEG metrics in different clinical delirium phenotypes, would like to discuss statistical methods to compare multiple metrics between different phenotype groups and then adjust for comorbidities

2024 July 23

Haatef Pourmotabbed, Biomedical Engineering

I am planning on submitting a proposal for the NIH F31 training program. My mentor, Dr. Catie Chang, will be present during the meeting.

We plan to employ multimodal neuroimaging, electrophysiology, and peripheral physiological data to investigated impaired autonomic brain networks underlying cardiorespiratory dysfunction in epilepsy.

In Aim 1, we will compare functional connectivity measures in fMRI between three groups (epilepsy patients with tonic-clonic seizures, epilepsy patients without tonic-clonic seizures, and controls). For Aim 1, we wanted direction on performing multiple comparisons correction for the statistical analysis (general linear model) and guidance on performing a power analysis for sample size justification.

In Aim 2, we will analyze functional connectivity measure in intracranial EEG data (stereo-EEG) collected during seizures in epilepsy patients. We will use a linear mixed-effects model to compare the functional connectivity measures between two seizure types (seizures with and without apnea). We wanted guidance on performing a power analysis for sample size justification for the linear mixed-effects model.

2024 July 9

Mariam Saad (Isaac Manzanera Esteve), Plastic Surgery

We have previously received funding from VICTR to conduct an MRI study in patients with headache, to visualize their grater occpital nerve and detect pathologies. This study is completed and we have an accepted publication of its results. We are now at a stage where we would like to refine the MRI coils used to image the greater occpital nerve, and we would like to compare resolutions and clarity of images received using a surface coil and a volume coil. We would like to discuss the sample size and the N in each cohort.

2024 March 26

Neil Dani, Cell and Developmental Biology

We are interested in characterizing the neurodegenerative processes at play in the choroid plexus, a region not routinely investigated in age-associated cognitive decline. To this end, we are collecting human autopsied tissue that will be available as and when donor cases come through the door, roughly 15-20 a year, across a range of ages and disease states. We seek advice on how to statistically interpret histopathology, sequencing data etc. from what is likely too sparse a dataset. We look forward to our appointment.

2024 March 19

Eric Wright, PMI

I am examining RNAseq data for Gene Ontology term enrichment by GSEA. I\x{fffd}\x{20ac}\x{2122}d like some assistance in interpreting the resulting statistics from this analysis.

2023 December 12

Nicole Malofsky, Biomedical Engineering

Determining TB bacteria drug susceptibility is essential to match TB-infected patients with their most effective drug regimen. In this work, we are developing an approach for confirmation of drug susceptibility without requiring sequencing. Melt temperature (Tm) of PCR products offers a rapid technique for confirming susceptibility down to a single nucleotide polymorphism using basic PCR instruments. A drug-resistant PCR product will produce a lower Tm than a drug-susceptible PCR product. In our application of isoniazid drug resistance, the mutations induce changes in Tm as small as 0.18\x{fffd}C. These small changes in Tm are difficult to detect on current calibrated PCR instrumentation due to instrument heating variability across wells and potential salt concentration variability amongst samples.

Our reagent-based approach adds an additional reaction component to help detect these small Tm differences. This additive, L-DNA, is designed to produce a reliable Tm representative of a drug-susceptible PCR product. In every sample, we can then compare this L-DNA Tm to the PCR product Tm to identify if the PCR product was drug-susceptible or not. Notably, L-DNA melting behavior is similar to that of a PCR product so these two Tm values will be proportionately affected by variance in heating. We are currently seeking to statistically assess this assumption.

To test this assumption, we used a 96-well format. It is known that current instruments have inherent heating variability across the 96 wells. A recent experimental design placed 24 drug-susceptible samples in the top left quadrant of the 96-well plate and 24 drug-resistant samples in the top right quadrant of the 96-well plate. The data was first analyzed by comparing Tm between PCR products of the two quadrants. An unpaired t-test of equal variance produced a p-value of 0.55. We interpret this to mean the \x{201c}gold-standard\x{201d} method cannot identify the difference between our drug-susceptible and drug-resistant samples. The data was next analyzed by subtracting L-DNA Tm from PCR product Tm. These Tm difference values were compared between the two quadrants using an unpaired t-test of equal variance and produced a p-value of 2.58E-05. We interpret this to mean the L-DNA method is indeed able to identify the difference between our drug-susceptible and drug-resistant samples. Is our experimental design and statistical analysis appropriate to draw these conclusions?

Next, we seek to assess whether our L-DNA additive can overcome variability in salt concentration amongst samples. This seems more complicated to test. All measurements affected by salt additives are also influenced by instrument heating variability. This adds a confounding variable to our testing.

To test this assumption, we used a 96-well format. The experimental design placed 18 drug-susceptible samples in the top left quadrant of the 96-well plate and 18 drug-susceptible samples with salt additive in the bottom left quadrant of the 96-well plate. We first wanted to assess whether Tm would be affected by accidental salt additions during variable reaction prep, for example. The data was analyzed by comparing Tm between PCR products of the two quadrants. An unpaired t-test of equal variance produced a p-value of 7.90E-23. We interpret this to mean the addition of salt significantly changes our PCR product Tm. We then wanted to assess whether the Tm difference values were consistent with the addition of salt. The data was analyzed by subtracting L-DNA Tm from PCR product Tm. These Tm difference values were compared between the two quadrants using an unpaired t-test of equal variance and produced a p-value of 0.053. We interpret this to mean the Tm difference values are not different when salt concentration is increased in samples. This would indicate our L-DNA comparison approach is not sensitive to changes in salt. Is our experimental design and statistical analysis appropriate to draw these conclusions?

2023 November 28

Jennifer Choe, Hematology/Oncology

I did a project with a pharma company that generated a raw data file of transcriptomic expression data in head and neck cancers. But I don\x{2019}t know how to analyze the data myself \x{2013}particularly in wanting to evaluate certain signaling pathways.

2023 November 7

Eric Wright, BME/PMI

I would like assistance in reviewing my RNAseq pipeline to identify problems.

2023 October 24

Uday Suresh, Department of Biomedical Information

Looking at audit log data and seeking help modeling behaviors based on aggregated observed measures.

2023 October 17

Eric Wright, BME/PMI

RNAseq on mouse liver; see good separation between groups by PC1, but very few genes with and adjusted p-value below 0.05

2023 September 26

Carl Stone (Megan Behringer), Biological Sciences

I grew a barcoded E. coli library over a few days and sequenced each timepoint to determine barcode frequency. Each library was grown in the presence of three different strains and every combination of the three. I am trying to measure barcode fitness in the presence of different combinations of strains over time and use Bayesian inference to determine how strain partners influence barcode fitness. I have been working on a Bayesian GLMM to accomplish this (using R packages ALDEx2 to model barcode measurement error and log-ratio transform samples and brms for the GLMM). My goals for the session are 1) to determine if a GLMM is the best method for this question; 2) to check that my data normalization methods are appropriate; and 3) to construct a model that fits my data and experimental design.

August 29, 2023

Valeria M Reyes Ruiz and Juan Barraza, Pathology, Microbiology, and Immunology

We have engineered reporter constructs for each of the 16 Two component signaling systems in S. aureus. Through the application of advanced imaging modalities, we identified three S. aureus TCSs that are activated upon sensing host-imposed stresses inside macrophages. We then developed an arrayed genome-wide CRISPR screen in macrophages with the use of robotics and high-content imaging to discover host factors responsible for the control of the intracellular reservoir of S. aureus and for the activation of these regulatory systems. Our data analysis pipeline relies on the use of machine learning with robust segmentation to measure both TCS-dependent reporter signal and constitutive bacterial signal on a per-cell basis for every well in our arrayed library. We are looking for help with the statistics on this dataset. We would like to identify hits and calculate confidence intervals for those hits. We have an idea of the data analysis, but would like to make sure everything is done appropriately with an expert on biostatistics.

2023 July 25

Mariah Caballero (Yolanda J. McDonald), Human and Organizational Development

Methodology review for data and spatiotemporal analysis strategy. Project entails community impacts (social and environmental dimensions) over nuclear energy development phases. Unit of analysis is census tracts within a 10-mile radius of nuclear power plant location. We are looking at three distinct phases of operational development, including installation, operations & maintenance, and decommissioning.

2023 April 18

Julie Bejoy, Nephrology and Hypertension

M2 macrophage exosomes to treat complications of diabetic kidney disease.

2023 March 28

Julie Bejoy, Nephrology and Hypertension

We focus our VICTR project on using M2 macrophage-derived exosomes on alleviating the complications associated with diabetes in kidney organoids. We will also do miRNA profiling of M2\x{fffd}-Exos by microRNA(mRNA) sequencing. The study will be performed on three or more samples on three or more independent occasions. We need guidance on statistical analysis plan for the grant application.

2023 February 14

Justin Jacobse (Yash Choksi), GI-Medicine

The aim of this analysis is to analyze cell abundance. The four independent variables (patient groups) are categorical. The 57 dependent variables (each variable is a cell type) are continuous and arbitrary units. There are 94 independent measurements per cell type (these are the 94 patients in the four groups) and there is no missing data.

The question is how to best determine whether cell frequencies are different between the patient groups. Whether one cell frequency is different than another cell frequency is not relevant. I have done a PERMANOVA with post-hoc SIMPER analysis and would like to verify this would be the correct analysis. The main reason to check is that it has been suggested that interpretation of SIMPER analysis is complex. Mentor confirmed.

2023 January 31

Nancy Newlin (Kurt Schilling), Computer Science

We want to explore how various pipeline parameters/decisions affect the reproducibility, repeatability, and bias of diffusion metrics (graph measures calculated from connectomes, bundle mean/length, FA/MD [all single scalar values]). We want to see specifically how changing one part of the pipeline (the number of streamlines/data points used during tractography) affects these things. We hope to define a certain number of streamlines/data points at which the pipeline becomes reproducible/repeatable/unbiased. We would like to know what tests we can use to determine reproducibly, repeatability and bias. Mentor confirmed.

2022 November 1

Carl Stone (Megan Behringer), Biological Sciences

We sequenced E. coli using nanopore sequencing to measure adenine methylation across the genome. This gives us the fraction of reads methylated or unmethylated across 38,000 sites in the genome, and we are comparing methylation at specific sites and genome-wide between 11 samples (3 groups). We are trying to 1) identify which samples are significantly different and by how much and 2) identify which sites differ between samples. I've done PCA but the large number of dimensions, weird distribution of the data (median methylation percent of ~97% but highly skewed), and multicollinearity between sites makes it hard to interpret, so I am looking for guidance on PCA or other methods to answer my questions. Mentor confirmed.

2022 October 18

Monica Morales (Dolly Ann Padovani-Claudio), Vanderbilt Eye Institute

Data was collected to analyze the inflammatory response of cells (N=3) after stimulation with 4 molecules (IL1b, TNFa, IL8, IL6) compared with vehicle. Qt-PCR gene expression of the same molecules was made and fold change was analyzed by one-way ANOVA. Questions: 1. p values in some treatment groups are significant and in other are not; even when the fold change is 5 times grater. Why? 2. When different experiments are compared. Some ANOVAS have significant p values and other don't. Even when they have the same fold change distribution in between treatment groups. Why? 3. Is there other statistical analysis that would work better for our study? Mentor confirmed.

2022 October 11

Shannon Townsend (Maureen Gannon), Molecular Physiology & Biophysics

Project deals with examining gene expression by qRT-PCR in cultured cells after different tretments singly and in combination.. Our issue is figuring out the best way to graph the data, as baseline measurements between treatments differs between groups. Mentor confirmed.

2022 23 August

Melissa Kimlinger (Josh Billings), Anesthesiology

My mentor and I have data from bulk RNA-seq in mouse kidneys that were adminstered different oxygen treatments during a renal ischemia/reperfusion surgery. We have done some analyses on our own, including PCA analysis, heatmap/unsupervised clustering, and some pairwise comparisons

to find differentially expressed genes and pathway analysis. As we are not statisticians, we would greatly appreciate expert input on next steps such as data presentation, statistical cutoffs, and recommended packages for pathway analysis. Mentor confirmed.

Justin Jacobse (Yash Choksi), GI-Medicine

Eosinophilic esophagitis is a food-mediated disease characterized by tissue eosinophils. We have a dataset of pediatric and adult tissue RNA-seq of esophagus. This data includes both patients and controls, and females and males. The dataset we use is composed of existing data as well as unpublished data from our lab. I use limma to analyze the RNA-seq data and have a question regarding the definition of contrasts; specifically how to examine/specify contrasts if I want to assess the effect of gender on gene expression in disease (compared to no disease) separately for adults and peds in one model, i.e. without two separate models. Mentor confirmed.

2022 June 14

Ciara Shaver, Medicine/Pulmonary

I have LC-MS/MS proteomic data from airspace fluid of organ donors. Samples are from about 15 patients and many have multiple timepoints. I need help analyzing the data into pathways most different between certain group and determining what proteins change most over time.

2021 September 21

Amanda Martinez-Lincoln, Alexandra Key, Special Education

Background: Proposal will examine shared reading in children and parents using hyperscanning. Specifically, the project will examine brain synchrony in parent-child dyads. Two groups: English Learners, Native English speakers. Two conditions: English-only text, bilingual text. We would like to address the justification of sample size and statistical analysis plan.

Meeting Notes: Received previous feedback to work on sample size and statistical analysis plan. The current budget only allows for 30 dyads (15 per year per group, 30 per group across two years). Aim 1: whether comprehension outcomes are moderated by language background (EL versus NE). Aim 2: EL only, whether comprehension outcomes are moderated by text type (English versus bilingual). Wanted to compare between conditions.

Recommendations:

1. Find the mean and standard deviation for comprehension outcomes and brain synchrony (both Aim 1 and Aim 2) from similar studies. Could use multiple studies with similar comprehension outcomes and similar population. Share the information with Hakmook Kang

so he can make a power analysis. No power analysis for Aim 3. In the SAP, mention that the study will be under-powered.
2. To calculate circular correlation in MATLAB, first take the Hilbert transformation (`hilbert`). This function takes the data and outputs the transformed data. After getting the output, can do `angle` in MATLAB. Then do `circ_corrcc` (within Circular Statistics Toolbox). This function takes the (output from `angle` for one participant, output from `angle` for the other participant) and gives the correlation. Do this on each combination of the dyads. Yan can send over the code. Alexandra Key will provide the reference.
3. Calculate the circular correlation separately for passive baseline, active baseline, and active conditions (two book reading conditions). Compare active condition separately with passive baseline, and with active baseline

2018 January 23 (Attending statisticians: Chang Yu, Hakmook Kang, Li Wang)

Joshua David Chew, Pediatric Cardiology Fellow

- The project is looking at pulmonary pulse transit time (pPTT), a novel marker for pediatric pulmonary arterial hypertension. There were N=21 PAH subjects with N=42 matched controls on age and sex.
- Two blinded reviewers completed the measurements on all of the studies. We reviewed the results of the ICC calculation and the Bland-Altman plot.
- We suggested fit a Cox proportional hazard model for the outcome.
- $5000 VICTR voucher would be sufficient for the scope of the work.

Sarah Fuchs, Pediatric Cardiology Clinical Fellow

- Pediatric cardiology patients were classified into two groups, those with in utero findings of RAS at < 30 weeks gestation (group A, N=15) versus those with in utero findings of RAS at >= 30 weeks gestation (group B, N=5).
- The primary outcome is Postnatal Transplant-free Survival comparing between the two groups. Since the event rate was low (8 in group A and 1 in group B had transplant or death), we suggest report Kaplan-Meier curve and the log-rank test for this pilot data and use it to plan for future study. Can also consider fit a Cox proportional hazard model using the actual gestational weeks of the RAS findings.

21 February, 2017

Joshua Beckman

I am a translational researcher who now has metabolomic data. This is a flux study wherein we have blood at baseline, after 5 minute ischemic stimulus and 1 minute after ischemic stimulus in healthy and diabetic subjects. I want to see if there is a difference at baseline, with the intervention, and how these baseline and flux changes associate with vascular function

14 February, 2017

Alissa Guarnaccia

I am a graduate student in Bill Tansey\x{2019}s lab in Cell and Developmental biology.

I have three sets of proteomic datasets (each ~1500 data points) that I am trying to perform the appropriate statistics for. Each dataset individually has been analyzed statistically, but I need help doing statistics on an average of the three. I believe I need to perform a Benjamini-Hochberg false discovery rate analysis to generate p-values of each datapoint. Ultimately I\x{2019}d like to generate a volcano plot.

High dimensional data is the theme my questions fall under.

31 January, 2017

Sophie Katz

I am a first year pediatric infectious diseases fellow and am in the very beginning stages of a project that would look at using a serum biomarker (procalcitonin) to aid antimicrobial stewardship efforts to pull off antibiotics earlier in the pediatric ICU settings. I'd love help with determining power and sample size, as well as any other advice you may be able to provide starting out.

20 December, 2016

Yuanjun Guo

I\x{2019}m Yuanjun in Hind Lal\x{2019}s lab in Pharmacology. I\x{2019}d like to consult about one of our PCR array study design. I prefer to come at this Friday noon (12/16) if there is still place available. The general idea of our study is we found the phenotype markers are significantly changed when we overexpress the gene X we interest under both control and drug-treated

condition. And we plan to move forward to find the target genes related with our gene and drug manipulation.

The **experimental groups** are harvest from cells as listed

Treatment	control	control	Treated	Treated
Cells	WT cells	Overexpression	WT cells	Overexpression

The **goal** of this study is to find target genes(those have significantly changed due to either overexpression of the protein VS WT in control and treated groups)

The **method** we plan to run qRT-PCR using a 384-well predesign PCR plates (96*4, for each 96-well part, there are **91** different target primers)

My main problems are what the minimal or proper replications I need for this study and if the sample size here will affect the **cut-off** later to pick up the genes significantly changed due to manipulation.

So far my plan is to have 3-4 different biological repeats of all four groups. I\x{2019}m not sure if this is enough and if later on just compare each gene between groups with unpaired t-test to find target genes. Also I\x{2019}m not sure if I need more than one sample per group and repeats for the same samples. Feel free to let me know if you need more information or any comments on this. Thank you so much.

22 November, 2016

Sarah Osmundson

I have analyzed my data for a project but need assistance deciding whether/how to make cut points and how to interpret the data. These are data from a prospective cohort study examining opioid use after hospital discharge. Most women use much less than prescribed but I am trying to make sense of women who use more than prescribed. Unfortunately due to my clinical schedule I cannot come on Mondays or Wednesdays.

Benjamin J. Reisman

David (cc\x{2019}d) and I would like to attend the 11/22 biostats clinic to review our flow cytometry data and discuss the best way to generate \x{201c}flow-cytometry like\x{201d} data to test out some of our analysis scripts and optimize our experimental set-up.

The problem we\x{2019}re trying to solve is that cell staining appears to be dependent on cell size, as reflected in FSC/SSC. We\x{2019}re attempting to barcode cells by staining cells form each well with a different level of fluorescent dye, which allows us to combine the wells, run them together, then resolve which well they originated from our experiment (ref: https://www.ncbi.nlm.nih.gov/pubmed/21207359)

Due to the correlation between cell size and staining, the largest cells in a low level may be brighter than the smallest cell in a higher level. I\x{2019}d like to optimize the concentrations of dye we\x{2019}re using modeling, but I need to generate model datasets to test on. If I have an experiment with cells of known FSC/SSC and Dye 1, how can I generate data with similar distributions covariances? Thought it might be most straightforward to subsample from our real data but was looking for expert input on how to best go about this.

25 October, 2016

Aaron Lim

Topic: Analyzing RealTime PCR data from multiple plates

Name: Aaron Lim

Affiliation: Vanderbilt Medical Scientist Training Program

Mentor: W. Kim Rathmell, M.D., Ph.D.

Dept: Medicine

I would like to attend the Tuesday, Oct. 25 clinic to discuss how to analyze RealTime PCR data from multiple plates, and how to compare gene expression from multiple tumor samples.

Carolina Pinzon

I would like to attend tomorrow again to double check some calculations. Thanks

11 October, 2016

Carolina Pinzon-guzman

I would like to attend the clinic on Tuesday Oct 11.

I am planning on running a proteomic study on amniotic fluid. This assay has never been done before on amniotic fluid in the second trimester of pregnancy. I am applying for VICTR funds and they are asking me for a sample size justification. I am planning on running 6 samples (3 control and 3 test)

Thanks

27 September, 2016

J. Brennan McNeil

Tomorrow, Tuesday September 12th, I am planning to attend the Omics Data Clinic to get some help with some mass spectrometry data that I have. I have two sets of data from two similar groups of patients which I would like to compare to one another. I have had some trouble doing this and my analysis is complicated by a few different things. Will there be a statistician present who can help me?

13 September, 2016

Shambnam Sarker

I would like to reserve a spot on Tuesday 9/13/16 noon to discuss my research which is a diagnostic cross-sectional study that I soon begin data collection on and appropriate analysis.

Consultants: Yaomin Xu, Alex Zhao

23 August, 2016

Zachary Dubit

Ayan Mukhopadhyay

I am a third year PhD student in the department of Computer Science, working in the Computational Economics Research Lab. My advisor is Prof. Yevgeniy Vorobeychik. My research interest is primarily Machine Learning, and I specifically work on predicting crimes and other emergency events. I need some help in understanding and answering a few questions about Survival Analysis. Tuesday looks like the best fit regarding this but I can attend on other days as well.

16 August, 2016

Zachary Dubit

I was wondering if it would be possible to attend the clinic on Tuesday, August 16th at noon in MCN for assistance with a statistics project using R and binary logistic regression. Thanks for your help and for offering this resource!

Sarah Kleiman

2 August, 2016

Matthew McKenna

Consultants: Hakmook Kang

21 June, 2016

Heidi Silver

Consultants: Hakmook Kang

I'd like to use the IDIOM data to look at changes over time in glucose, insulin and c-peptide, and the various indices of insulin sensitivity and resistance.

10 May, 2016

Dara L. Eckerle Mize

I am hoping to get a little help with multiple imputation in R. I am using aregImpute on a small dataset (10,000 observations of 2 predictors) but I am not completely sure if I am doing it correctly. Please let me know if this is not something that can be discussed.

Benjamin K. Poulose, Associate Professor of Surgery

(at 1pm) Sampling methods to perform long term follow-up in a surgical registry.

26 April, 2016

Julian Peters

I need some help on diagnosis and prognosis data that I have collected. This data is from a longitudinal study collected at various time points during treatment for each patient. I would like to create trends in the data by each diagnostic method and then to determine if the 3 diagnostic methods are predictive of each other, if they correlate and to what extent they correlate. I would also want to determine if these diagnostic methods are predictive of treatment outcome.

12 April, 2016

Lucy Spalluto

I am working on a faculty development project. We are assessing the utility of educational modules. An anonymous pre-module survey with yes/no questions is sent to faculty members. An educational module is presented (not all faculty attend, material is made available online for review to everyone). An anonymous post-module survey with the same set of yes/no questions is sent. Two additional questions ask if they attended the event and whether or not they reviewed the educational materials.

28 March, 2016

Eric Rellinger

Consultants: Yaomin Xu

I am a research fellow working in the laboratory of Dr. Dai Chung. We have been collaborating with Dr. Beauchamp on evaluating the effects of his compound ML327 on neuroblastoma growth. We have had a pretty robust phenotypic response and are planning to perform RNA sequencing of neuroblastoma cells treated at early and late time point to 1) identify potential upstream regulators, 2) identify potential compensatory pathways that mediate cell survival in the presence of ML327, and 3) determine whether the fate of these cells are altered in the presence of the compound (i.e. differentiation). I reached out to Dr. Zhu and she thought that you might be a useful resource for biostatistical support for this endeavor.

I have written a VICTR grant to help support the funding of this endeavor and was curious if you would be willing to aid with the biostatistical interpretation of those results. VICTR submissions require a biostatistician before submitting the proposal. In total, we are planning to have 8 different groups which will be completed as biologic triplicates, resulting in a total of 24 samples submitted for 30M 75bp paired end reads.

Any thoughts or feedback would be greatly appreciated. I would be happy to meet as well if you wanted to discuss the project and its goals in further detail.

Best wishes,

Eric

Comments:

1. Independent cell lines will be used for measurements at two time points.

2. Suggest to identify genes based on foldchanges comparing each pairs of conditions and use 2 foldchange as a cutoff. Given relatively robust phenotypic response and expacted effect size, three replicates at each experiemntal conditons (24 samples in total) should provide sufficient power in this discovery project to generate preliminary results.

Flavio Silva

My name is Flavio Silva and I am a physical therapist with the Department of Orthopedics. I just finished data collection for a case control study about Injury prevention for musicians: SCAPULAR AND CERVICAL NEUROMUSCULAR DEFICITS IN MUSICIANS WITH AND WITHOUT PLAYING RELATED MUSCULOSKELETAL DISORDERS: A CASE-CONTROL STUDY

IRB NUMBER: 141569

I was wondering if I could get some assistance with the analysis. The project is through the department of orthopedics. The data set is very simple but I need some help with running a regression and with descriptive data set.

02 February, 2016

Kimberly Albert, Vanderbilt University

I would like to come to a walk-in clinic for help in planning the analytical plan for the study described below for a grant application. I am happy to come on a different day if Thursday is not the appropriate topic.

Age-matched older men and women will be compared. Cognitive performance will be quantitatively assessed at screening and following a psychosocial stress task. A one year follow-up assessment will be completed to examine the predictive value of neural activity during psychosocial stress on long-term cognitive status.

Aim 1: Examine sex differences in psychosocial stress induced changes in acute cognitive performance and functional connectivity in older adults with Subjective Cognitive Decline (SCD).

Aim 2: Examine whether neural activity during psychosocial stress in older adults with SCD correlates with long-term objective cognitive performance change over one year.

19 January, 2016

Akshitkumar Mistry, MD, Vanderbilt University

I would like to review meta-analysis techniques. I have conducted a meta-analysis correlating brain tumor location with survival. However, survival in brain tumor is also dependent on other variables such as age, extent of surgical tumor resection, etc. My meta-analysis shows a statistical signficant effect of brain tumor location with survival; however, I do not know how to account for the confounding variables (age, extent of resection, etc...). I would like to show you the data and review statistical techniques so that I can apply for a VICTR voucher.

Charles Caskey, PhD, Vanderbilt University

Discuss collaborating with someone on an upcoming grant submission

12 January, 2016

Sandeep Arora, MBBS, Department of Radiology and Radiological Sciences, Vanderbilt University

- I was hoping to attend a clinic to discuss a grant submission for the following project - submission deadline is Jan 15 (next tuesday will be great). Abstract - attached. I need to discuss power and statistical analyses methods.

- Phase-Shift Nanodroplet Assisted Multifocus MR guided Focused Ultrasound Ablation of Lobar Portal Venous Supply and Ipsilateral VX2 Hepatic Tumor Implants in a Rabbit Model.

15 December, 2015

Karthik Sundaram, Vanderbilt University

- I'm currently at Vanderbilt Radiology Resident looking to work on an imaging project related to imaging prostate cancer in patients undergoing a prostatectomy and we could use the biostat department's help in calculating costs.

- Briefly, we plan to image patients with prostate cancer that over-expression of a receptor called the organic anion transporter. Evidence suggests that 50% of prostate cancer patients over-express the receptor. We plan on staining for the receptor post resection. Based on this information, we would like the clinic's help in calculating the number of patients in our pilot study that we need to image in order to have success in our study.

08 December, 2015

Chan Gao, MD. Ph.D., Physical Medicine & Rehabilitation

Consultants: Yaomin Xu, Run Fan

I will undertake a research project "genome-wide association study of rotator cuff tear". I am submitting application for BioVU access. The stat analysis method needs to be described:

Briefly describe the method for and statistical analysis.

Include sample size estimation, dependent/outcome variable(s), independent variables (include SNPs, covariates, confounders), type of statistical model (if appropriate), how SNPs will be coded, power calculation/ population stratification plans.

- Gave a very detailed overview about BioVU. Refer to Dr. Todd Edwards for more support.

15 September, 2015

Matthew Rioth, Vanderbilt Ingram Cancer Center

Consultants: Yaomin Xu

- I am proposing an investigation of the association of observed tumor genetic variants and time on treatment phenotype. Specifically, we have a database of variants in ~800 tumor samples detected by exome panel sequencing (~300 genes, median 12 variants per sample). We have the ability to extract treatment information from the EHR. We would like to correlate variants, grouped by protein functional region (ie ligand binding domain, DNA binding domain, catalytic site, etc) with treatment response.

- A hypothetical discovery could be that patients with mutations in a linker region in the estrogen receptor gene have much shorter time on treatment with fulvestrant (an estrogen receptor antagonist) relative to mutations in other regions of the gene. This implicating mutations in that region as a mechanism of fulvestrant resistance. Since we will be able to extract time on treatment information from the EHR, the analysis we were proposing would be a Cox (time to event) regression. Does this sound appropriate? Is there a lower limit or guidance for how few patients to have in a group in this analysis? If we were to perform this analysis for multiple phenotypes (hypothetically 10-20 phenotypes) associated with multiple genotypes, what kind of multiple hypothesis correction should we use?

30 June 15

Ping Wang, VUIIS

5 May 15

George Nelson, MD Assistant Professor of Medicine, Division of Infectious Diseases

Consultants: Fei Ye, Pengcheng Lu.

- * I have a quick question about power calculations for a non-inferiority cross over trial. I will bring all relevant numbers
- Rate of MRSA Acquisition data: This would be a cluster randomized cross over trial with standard of care being isolation of those with any MRSA isolate and intervention being isolation of only draining MRSA wounds. The intervention period would be 3 months with 2 week wash out and cross over to standard care (control)/intervention status. Literature reports 3.5-5/1000 patient days acuisition of MRSA in clinical care. WE have ~ 6 units available for study with 15000 patient days over 3 month period.

14 April 15

Jack Virostko, Research Assistant Professor, VUIIS

- retrospective matched case-control study of pancreas imaging in type 1 diabetes

7 April 15

Gabrielle Rushing, student, Neuroscience Graduate Program

Consultants: Fei Ye, Hakmook Kang, Pengcheng Lu.

- Analysis of flow cytometry data: expression of different phospho-protein levels in neural stem cells
- Compare region to region variations and variaion between three mouse strains.
- Paired comparisons and multiple group analysis: Wilcoxon signed rank test (paired) and Kruskal Wallis test (multiple groups)
- More complicated correlated data analysis: mixed effects model with appropriate number of samples.

17 March 15

Akshitkumar Mistry, resident, neurosurgery

- Retrospective study of 57 patients who underwent surgery for trigeminal neuralgia on one side of brain and whose MRIs on both sides were later examined. Two factors were examined: (1) degree of contract (none, contact, compression) (2) location of vascular contact (at REZ and distal to REZ).
- Hypothesis: location and degree of contact are associated with the side of surgery (each patients had information on two sides and only had surgery on one side). Mixed-effects logitstic regression model on surgery (yes/no) on the two factors.
- Hypothesis: location and degree of contract are associated with early outcome (0, 1, 2, 3) on surgical side. Ordinal logistic regression model on the two factors.
- Suggest applying for $4000 Voucher to perform the statistical analysis and prepare the manuscript.

10 March 15

Attendants: Poojitha Matta, and Mentor: Stacy Sherrod, PhD Department of Chemistry.

Consultants: Fei Ye, Steven Chen, Pengcheng Lu.

- My project aims to characterize the impact of chorioamnionitis (a maternal infection contracted during pregnancy) on the fetal immune system.
- Two group of different patient samples, one is Chorio + and Chorio - (i.e. infection before birth in mothers), 10 samples in each group. Within each group, the patients were in two categories: one was control, the other was activated by Acid B and CD28. There were ~1600 features with normalized metabolites intensities.

- Sample size is too small. The analysis will be under-power. Two factors, both were binary variables, and there were correlation issues as well due to the measurements on same sample.
- The simplified linear regression model: Activated ~ Control + Chorio + error term.
- Another naive way: Use the difference between Activated and Control as outcome measurement, compare the difference between two populations. Don't report p values.

20 January 15

Attendants: Antonio Hernandez, MD, Liming Luan, PhD, and Edward Sherwood (Mentor), MD/PhD Department of Anesthesiology.

Consultants: Fei Ye, Steven Chen, Yaomin Xu, Pengcheng Lu and Li Wang

- Our lab has been involved in a study of challenging human neutrophils with either lipopolysaccharide or Monophosphoryl Lipid A, for the evaluation of inflammation. We looked at gene expression for inflammation. We have the data, but we are unclear as to how to proceed. Specifically, how to analyze the data and whom to approach for assistance.
- Experimental Design: 18 people's blood samples. Six samples with PBS, six samples stimulated with LPS, and six samples stimulated with MPLA (monophosphoryl lipid A)
- Suggestion: DE analysis, Gene pathway analysis to cluster genes, and gene set analysis (GSEA).
- We prefer getting the raw data, .CEL files, to start the processing to assure the quality of normalization.
- Contact Steven Chen for further collaboration.

13 January 15

Attendants: Carolina Pinzon-Guzman, MD/PhD

- I'm a PGY2 surgical resident writing an IRB and VICTRL proposal but I am stuck on the statistical part trying to figure out how may patients I need for my study. I would like to talk to somebody about it. I can go to the clinic tomorrow, or I can meet somewhere else. It is a simple question. I am looking at the amniotic fluid in pregnant females with babies with congenital abnormalities and comparing the amount of some growth factors in it vs control amniotic fluid. Nobody has ever done this experiment and I am not sure we would be able to find a difference.
- How do you predict how many amniotic fluid samples I need?
- What if I want a do discovery proteomic study looking at a difference in protein amounts in the amniotic fluid?
- Note: Showed up at 12:50pm, she decided to come to tomorrow's clinic.

Attendants[Stop by client]: Manisha Gupte, MD. Cardiology

Consultants: Dan Ayers, Hakmook Kang, Pengcheng Lu, Alex Zhao

- Question: A Pilot study with four ECHO treated samples and three untreated/control samples, outcome is continuous measurement of activity at week 0 (baseline), week 1 through 7. Any suggestion on group comparison for treatment effect?
- Not statical testing of comparison is need ed for pilot study especially when sample size is too small.
- Try to find the minimum biological difference you would like to detected.
- Define your time points, why did the experiment end at week 7 rather than week 9?
- Compare the data at last time points make more sense than doing it at each time point.
- Need to adjust data for baseline.
- Contact Frank Harrell if statistician's help is needed in the future.

6 January 15

Attendants: Dr. Adrienne Dula

Consultants: Hakmook Kang, Li Wang

- I am an imaging scientist and am working on a VICTR application.
- Iron-based new contrast agent for liver imaging, enhancing T1 & suppressing T2 compared to current contrast agents. This can be metabolized by the liver.
- Diagnosis for patients with liver diseases & with chronic kidney diseases
- Current way for diagnosis is still non-invasive but not applicable to whom have both liver and chronic kidney diseases - Aim 1.
- Recruit patients with known liver cancer and no kidney disease: can get ferumoxytol (new agenets). Step2, search their previous MR images (within a month) with gadolinium (Gd) - based contrast agent (old, gold standard for now). Compare the new enhanced images with the corresponding standard images. Variables: size and number of tumor.
- Aim 2: Fibrosis ---> Cirrhosis: Invaisive method and not reliable. New MR-based imaging tool for non-invaisive and reliable quantification of the degree of fibrosis --> build prediction model at the end.
- $10,000 pilot VICTR pilot project (matched by John Gore, sum to $20,000)
- $267 ferumoxytol per person + $500 per hour for MRI ~ $850 ---> 23 subjects.
- Power analysis based on this number --> claim that this sample size is capped by the budget and maybe under-powered.
- $2000 (35hours) biostat voucher is recommended.

16 December 14

Attendants: Dr. Jane Ferguson

Consultants: Zheng-Zheng Tang, Guanhua Chen, Pengcheng Lu, Li Wang.

- I am hoping to get some advice on analyzing and integrating multiple high-throughput datasets, with the primary aim of developing a coherent analysis plan for a grant submission. Based on the information on the website, I believe that the Tuesday clinic for high-dimensional data may be the most appropriate. I work with human samples, and am proposing to integrate multiple levels of omics data in ~100 healthy humans. I will have data on gut and oral microbiome composition (sequencing data, collapsed into 100\x{2019}s of bacterial genera), gut and plasma metabolomics data (100\x{2019}s to 1000\x{2019}s of metabolites), plasma circulating microRNAs (100\x{2019}s), as well as dietary information. I am submitting this grant in January, and would very much value input on the statistical analysis strategy.
- Question: Whether diet will affect the microbiome and metabolomics data;
- Suggestion: First, lay out all the hypotheses, which one is the most important one? Enough for R01, better to limit a bit of the study aims;
- Aim 1: Association between microbiome and diet and health markers; Find common signatures or compositions between same person's gut and oral? May separate them and do not do the comparison at this point; Aim 2: association between microbiome composition, diet and metabolomics markers; Aim 3: Identify miRNA markers associated with diet and microbiome;
- Software and Statistical Methods: 1. The HuttenHower Lab developed a series of software. 2. Hongzhe Li's group at UPenn; 3. Can model the associations by considering microbiome as dependent variable (predictor variables: health markers, dietary data), and for dependent variable metabolomics (and lower level microRNA markers), treat microbiome as predictor variable.
- Contact faculty: Zheng-Zheng Tang and Guanhua Chen for grand proposal assistance.

11 November 14

Attendants: Drs. Lucy Spalluto & Rifat Wahab, Women's imaging

Consultants: Hakmook Kang, Steven Chen, Pengcheng Lu, Min Gao.

- Data: Questionnaire result/summary worksheet.
- Suggestion: Most of the questions can form a r x c table and be analyzed using Chi-squared test. Dr. Kang is the primary statistician to collaborate with the Department of Radiology.

Attendants: Anne Kenworthy and PostDoc Krish

- My postdoc Krish and I would like to attend the clinic this coming Friday (11/14) to discuss some data collected from microscopy images. (It is not high thruput information so I felt like

Friday would probably be more appropriate than Tues.) I am attaching an example of the data we are analyzing. We want to analyze the % of cells that contain tubules under different experimental conditions. We have also analyzed other tubule properties such as their length. We have some ideas about what statistical tests to apply this but would like asecond opinion to make sure we are doing this correctly.
- Suggestion: Poison regression model for count data by taking into account for the experimental conditions and offsets.

Attendants: Dara Mize

- I am working on a project for a Bioinformatics course and am attempting to perform some statistical analysis in R using 2 data sets from GEO. The data sets each contain microarrays for paired normal and papillary thyroid cancer tissues. I think I would like to compare gene expression between normal and cancerous tissue in each set individually. Then, I would like to analyze gene expression between normal and cancerous tissue when the two sets are combined. I would appreciate any feedback about this approach and have some specific questions about using R with these datasets.
- Suggestion: 1. Use Bioconductor package GEOquery to retrieve gene expression data; 2. Two data sets shared the same platform, but could not combine them directly due to different normalization methods and batch effect. 3. Paired and non-paired t-test could be applied to two data sets respectively. 4. Need to adjust nominal p values for meaningful interpretation (FDR).

30 September 14

Attendants: Christy Pearce, Asst. Prof. Division of Maternal Fetal Medicine, Department OB/GYN

Consultants: Hakmook Kang, Yaomin Xu, Steven Chen, Pengcheng Lu, Li Wang, Xue Han, Min Gao and a student from Dan's Biostatistics class

- I am not able to come to a Thursday clinic, so was advised to come to a different noon clinic. I would like to sign up for 9/30. I have run some stats in JMP, but would like these checked for accuracy as well as another calculation. Also, I would like help in the BioVU application on this particular SD dataset. Thank you in advance for your direction and help. Please let me know what else you may need.
- I have attached my data as well a pdf's of the JMP analyses I ran. I also attached the abstract that I submitted for our annual SMFM meeting. This data was from the SD, so it is de-identified. Goal was to identify risk factors for cardiac dysfunction (defined on echocardiography) in patients with preeclampsia. Design is case- control. Cases: preeclampsia + cardiac dysfunction, and controls: preeclampsia without cardiac dysfunction.

Exposure is BP, creatinine, LDH, uric acid, etc. In the end, really is more cross-sectional since the "exposure" typically occurs very close to the cardiac dysfunction.

- Questions:
    1. Are the analyses I ran accurate?
    2. Is there a regression analysis that can look at all the "exposures" and give a risk of cardiac dysfunction with additional "exposures" or way to make a scoring system based on exposures seen in each patient. So if a patient comes in with DBP >110 and CR >1 for example, her risk is _ as opposed to a woman who only has the elevated BP.
    3. Need help with sample size calculation to use this data set for GWAS data analysis as well as specific genes, including:
        1. PEE 1- chromosome 2p13;
        2. PEE 2- chromosome 2p25;
        3. PEE 3- chromosome 9p13;
        4. PEE 4 STOX1 gene chromosome 10q22;
        5. PEE 5 CORIN gene chromosome 4p12;
        6. EPHX chromosome 1q, EPOX, HYL1, MEH, EPHX1 (all known as epoxide hydrolase, microsomal [xenobiotic);
        7. NOS3 (ECNOS, eNOS, NOS3, nitric oxide synthase 3 [endothelial cell].
- Design: Based on the availability of the SD database, only 33 cases were obtained in this study, and 99 control samples were included(3-to-1 ratio).
- Suggestion and concerns: Regarding the JMP output, some concerns are: a few categorical variables need to be treated carefully with multiple levels and very small number issue, may combine some levels. Rethink lots of variables with large numbers (>75%) of missing values, could be excluded if they are not important in term of the research interest.
- Modeling strategies: Due to 33 cases only, three predictor variables can be included in the logistic regression model as a good statistical practice, but predictor variables could not be chosen based on the univariate analysis result (from JMP). If including variables based on literature of research interest, very likely there will be more than three independent variables, can think of the penalized logistic regression model, but the power is still a big concern of this study. Doing meta-analysis if possible by trying to find other similar research data sets outside to increase the sample size.
- It is not at the stage of sample size determination for SNP analysis yet, but obviously there is not enough case samples available in the SD db.
- Contact Dr. Chris Slaughter first to check if there is collaboration plan between him and the Department of OB. Since APS funding is available, fee-for-service model will also work possibly.

26 August 14

Attendants: Scott McCall, MD/PhD Candidate

Consultants: Yaomin Xu(contact), Pengcheng Lu, Meng Xu, Alex Zhao, Derek Smith and Allison Hainline

- My name Scott McCall and I'm a graduate student in Billy Hudson's lab. I'm currently working on a project with Josh Denny involving abdominal aortic aneurysms and BioVU and I'm at a statistical impasse! I would like to schedule a time this week if possible. Based on the descriptions of the different focus of each day, I'm unsure which would be best.

The root of the problem is that there is a seemingly strong, non-linear dependence on some covariates which I am having difficulty incorporating into the final regression analysis as part of my overarching story. After working with friends in Bioimformatics (Jacob VanHouten and Pedro Teixeira), we have been unable to solve my problem of trying to incorporate splines into some of my regressions to capture these associations in an internally consistent way. So it is my hope that we can work on this as part of the clinic. In addition to the data set (as per the specifications on the wiki), I am including the initial draft of the figures so we can hopefully fold this into the discussion also .

- Suggestion: fit restricted cubic splint regression model to catch the nonlinear effect of age, say let pulse pressure be the outcome; nonlinear interaction between age and G allele dose levels needs to be tested before drawing conclusion of "age dependence of SNP effect"; Race may also need to be adjusted in the model if the information is available, which may improve the precision of estimation of the allele effect.
- Follow up with Yaomin for potential statistical collaboration.

22 July 14

Attendant: Spyros Kalams, Vanderbilt HIV Vaccine Trials Unit

- I would like to request some time for a VICTR biostats studio Friday 6-26-14. Dr. David Aronoff and I are putting together a small study of 15 individuals. We are evaluating the effects of a drug on T cell activation, and we are comparing the effects of two different routes of delivery of the drug. I can go over this in more detail during the studio, but we would basically like to describe the way we will analyze the data in the study protocol.
- Primary outcome will be the T cell activation - percentage (continuous), which will be measured 4 times from day 0 to day 28. Descriptive statistics about data distribution like median and IQR can be calculated. Linear mixed effects model can be used to explore the activation change over time.
- Fisher's Z transformation or logit can be applied to make percentage goes from negative infinity to infinity.
- Count data can be analyzed using Poisson regression.
- If multiple biomarkers are processed simutaneously, p-values need to be adjusted.

Attendant: Aliya Gifford, Graduate Student (Chemical and Physical Biology)

- I'm a graduate student here and would like to attend a biostatistics clinic, and if possible I'd like to attend either Tuesday, Wednesday or Thursday this week. The questions I have are in regards to regression analysis, and understanding how and what approach to take, given the measurements we have. If there's more information I need to supply, please let me know.
- There are total N=18 subjects with brown fat and N=5 subjects with non-brown fat. Each subject underwent both CT and PET. There are also measures of FSF and R2s.
- Primary question: whether imaging results with BMI, Gender, Waist, Age, Height, Weight, and Temperature can be used to predict whether that patient has brown fat or not.
- logit (Y=brown fat) = CT + PET + FSF + R2s + demographics. Need penalization. Consider interaction between temperature and imaging.
- Current data is ROI based, can consider voxel based analysis.
- Scatter plot for all the possible pairs together with spearman correlation coefficient for descriptive analysis.

Attendant: Daniel Croymans, internal medicine

- Primary outcomes: number of hospitalization, job productivity, sick days, Dependent variables: BMI...

13 May 14

Attendants: Roop Gill, Plastic Surgery

Consultants: Bill Dupont

- Project looking at risk factors for rhinoplasty complications.
- There are two data bases. One is about cosmetic procedure (~18,000), and the other is the complication data (~2,500). Need to merge them together by some common variables.
- Will apply for a VICTR voucher and suggest $2000.

25 Feb 14

Attendants: Jennifer Herington and mentor Jeff Reese (not in attendance), Department of Pediatrics/Neonatology

Consultants: Pengcheng Lu

- Study objects: Identify novel inhibitors of intracellular calcium release from uterine smooth muscle cells.
- Method: Use high-throughput screening to measure changes in intracellular calcium release (indicated by relative fluorescent units) from mouse and human uterine smooth muscle cells (UT-SMCs). We will base our analysis off of a publication from Prudencio Dianas Vol 2 No 1 2013. First, plates contain compounds run in singlet only, for quality control and to determine the threshold for antagonist activity (selection of drugs). A z'factor and coefficient of variation will also be performed for quality control (robustness of assay to use in high-throughput screening). Once we determine "hit" compounds, we will repeat the assay 3 more times, therefore increasing our N-value to 3 but this will still be too small for a t-test.
- Suggestion: We will use linear models and moderated t-statistics as implemented in Limma in Bioconductor to overcome the small sample size issue.
- The estimated time to process HTS data is about 6 hours.

5 Nov 13

Attendants: Parimal Samir, Andy Link, Chris Browne, Ryan Delahanty and Rebecca Levinson

Consultants: Lily Wang, Steven Chen, Pengcheng Lu

1. Questions of Parimal Samir (from Andrew J. Link Lab: Dept. of Pathology, Microbiology and Immunology):
- Study objects: 1. Study the response of cells when its growth condition is changed from its happy physiology condition; 2. Find modules in the proteome that respond to specific changes in the physiology condition; 3. Find the combinatorial effect of the simutaneous changes in growth condition.
- Method: Compare 3 experiments and one control group, then classify the winner proteins(based on p values from ANOVA tests) into classes(based on FC>1.5 and p
- Suggestion: Fisher's exact test and Kolmogorov-Smirnov tests both assume all proteins are independent, this could be incorrect, can use GSEA. They updated the db recently and can input gene symbol as well.

1. Questions of Ryan Delahanty and Rebecca Levinson (from Epi / Human genetics division)
- Question: logistic regression to test the combined phenotype, genotypes.The outcome is disease/condition, there are 1600 different medical conditions, predictors are covariates like age, principle components and CNV(~1000 CNVs, coded as 0, 1, 2, NA). It's a kind of "phewas" study. The sample size is around 3000 patients. Would like to permute on CNVs, but the problem is when NA presented in CNVs, the outcome will change.
- Suggestion: Using FDR to adjust the p values from 1000 tests, you can use q-value as well; can also try exact logistic regression to get the exact p values.

20 Aug 13

Rockann Mosser, post-doc, and Maureen Gannon, Division of Endocrinology, Diabetes, and Metabolism, Dept. of Medicine

Dr. Dave Tabb told me to contact you (and ask for Ming Li if available) for some statistical help with some protein mass spec data sets I have. Basically, I have 2 studies: one with rat serum and one with mouse serum. The rat study compares 4 sets of rats (each set with an n of 3): 2 month old control, 2 month old treated, 6 month old control, and 6 month old treated. The mouse study compares 2 sets of mice again with an n of 3): control vs treated. I have quasitel.tsvfiles comparing age matched rats and the mouse set (so 3 comparisons, so far). I want to do more complex comparisons / statistics with these data sets, but I am not sure how best to proceed or even if the quasitelstatistics are correct for the comparisons I have (in other words, what I should be looking at).

13 Aug 13

Carrie Moore, Graduate Student, CHGR

To discuss penalized regression, particularly Lasso/Ridge/Elastic Net analyses. I have a couple of scripts in R which utilize glmnet and caret packages to perform elastic net analyses. In one of the scripts, some of the predictors which I know to be highly correlated with the outcome are not showing up in the final models. Is anyone in the biostatistics department with knowledge of elastic nets or those particular R packages available to answer a few questions?

- Consider the purpose of this model. Is it prediction or selecting relevant variables?
- Try running the procedures on your entire dataset and see how the result compares to those of your CV/bootstrap programs
- Consider using bootstrap internal validation
- Bootstrap and cross validation are useful for telling you how well your model performs, not for selecting variables.
- The major purpose of the cross validation is the variable tuning for the elastic net parameters (alpha and lambda).
- Have 180 individuals with about 90 events. Since there are 15k candidate predictors, the dimensionality is really too big.
- Sure independence screening: two stage method. First stage looks at correlations of the vars with the outcome. This uses cross validation. Stage two is something like lasso or elastic net.
- Random lasso. Discussed that the interpretation of the results would be difficult.
- Consider using ridge regression since it has only one tuning parameter, whose estimates will be more stable. To select the lambda, try bootstrapping the process. IF the lambda parameter does not vary much, then you might avoid the bootstrap. Then in the resulting model, could try "model approximation" where you fit are really non-parsimonious model and then try selecting a subset of the variables that represent about 90% of the signal. Can do this with a stepwise selection process that is blinded to the actual outcomes, but instead uses the predicted values from the model fit.
- see optimism bootstrap
- Consider forcing FEV, age, and height to be in model. Since you know they are important, consider using splines to allow them to behave nonlinearly.

Georgia Wiesner and Kelly Taylor, Department of Medicine

- genetic testing sometimes reveals actionable mutations and variants in genes that were not part of the reason for the testing.
- ACMG has a policy stating that there is a group of genes variants for which, if there is a discovery, the information should be passed to the clinician. Some of them are in cancer and some are in cardiovascular health. The question of interest is with identifying these variants in children.
- Planning a VICTR funding submission and looking for input on the statistical analysis plan.
- Part of the study's goal is to build an anonymized data set in biovu to assess the frequencies of deleterious mutations.
- Need to know the sample size.
- Calculate the sample size needed to achieve a given precision in the estimate of proportion. Find the maximum prevalence level you are likely to see, and base the calculation on the maximum. This will be the "worst" case in terms of precision.
- There is a five-level rating scale to categorize the prognosis for different variants, ranging from deleterious to uncertain to beneficial.
- Consider getting the frequency of deleterious variants separately by function or by population groups.

25 June 13

Cody Wenthur, Pharmacology : Peter Martin(Mentor, Psychiatry)

Statisticians: Pengcheng Lu, Yuwei Zhu

- Cody is a Ph.D. student of Dr. Craig Lindsley, and Dr Peter Martin.
- Project: Measurement of variation in metabotropic glutamate receptor-encoding genes (GRM3) amongst substance-dependent individuals: Implications for development of novel therapeutics
- Used CaTS softare to determine power, recommended use of PGA software with additive design, and matched study design
- Recommended attendance at VANGARD design studio for sample size verification and power calculation along with data analysis plan
- Roughly estimated hours for VICTR grant: 60 hours (analysis and manuscript preparation)

14 May 13

Katherine Betke, Pharmacology

Statisticians: Pengcheng Lu, Hakmook Kang

- Katherine is a Ph.D. student of Dr. Heidi Hamm. Her project is to compare pre-synaptic and post-synaptic protein expression enrichment data.
- G protein expression data of six mice was collected in both pre- and post regions, the null hypothesis is there's no difference of protein expression data between two regions.
- Check data distribution to see if log-transformation is needed; Can do paired t-test, but Wicoxon signed rank test is preferred.
- With log-transformation, the logFC will be the mean difference of expression data, zero means no change and FC is 1.
- Data visualization could be either scatter plot or bar chart. For bar plot, can add either error bar or 95% CI to show data variations.
- Will do for >20 proteins. Then multiple testing correction is needed. Bonferroni or FDR method could be used, but Bonferroni is too conservative.
- In R, p.adjust() can do p value adjustment by specifying methods.

7 May 13

Leslie Halpern, MMC

I have a project based upon previous work that I did. It is a diagnostic protocol using injury location and a screening questionnaire to diagnose victims of violence/abuse. I am well-published in this area. I am now involved in a project using this protocol with another well statistically used questionnaire and using salivary markers to measure inflammatory peptides in victims as compared with controls. I need to go over the use of my multivariate model to include the salivary sampling. I also need to discuss the power needed for this pilot study since finding is small and I need to but kits for the sampling. There are 3 other investigators in this project including: Johnson, Desiree; Gangula, Pandu R.

23 April 13 (probable)

Daniel Cohen, Pathology Resident

To discuss my project in anticipation for a VICTR voucher for a BioVU Exome Chip comparison study to identify SNPs for a patient population of interest. With this and other genes identified I will pursue NGS sequencing of patient cancer tissue and VICTR grant to fund in part the sequencing. Attached is an abstract relevant to the study. I have discussed this study with Dr. Yu Shyr. The introduction and services from your department were suggested from the VICTR office. I request your guidance with Methods 4 and 7 from the abstract: 4) BioVU Exome CHIP analysis of otherwise healthy non-BRAFi patients with cSCC (233 pts as of 2/2012)and age, gender, race matched healthy non-!cSCC controls to assess SNP association frequency with cSCC not present in normal individuals. Perhaps 699 BioVU patients could be extracted for control (3x)? 7) Biostatistics association analysis of lesion histology (verruca, AKs and BRAFi-cSCC of each morphologic subtype), HPV infection rate and host-genotype from NGS data of ~307 skin lesions from 55 patients (see table "...pt summary".

- Material has been placed in the clinic room main computer in the `clinic/hddata` folder

19 March 13

Stacy Sherrod, Physics and Astronomy Department

Consultants: Ming Li

Project: Metabolomics Mass Spectrometry Data

- The types of samples and data: Copied from Stacy's email:

What types of samples I have\x{2026}

1. Typically a control, load, quality controls and a few treatment types (in the data I have attached, I have control (C), load (LOAD), quality control (QC), astrocytes (A), microglia (M), neuron(N))
    1. Important: Loads and Quality Controls are the same sample, \x{2018}Load\x{2019} samples condition prior to running any other samples, typically 10 injections, the \x{2018}quality control\x{2019} samples are ran every 10 samples. Both Load and QC samples should group together\x{2026} and should account for instrument variability over the course of all the runs.
    2. Both \x{2018}Load\x{2019} and \x{2018}Quality Control\x{2019} samples are an equal mixture of all samples in the sample set.
    3. Sometimes I\x{2019}ll have a time course and different treatments.
    4. I try to do 3 biological replicates, BR) of each sample and 3 technical replicates of each, but don\x{2019}t always have enough sample for that, or misinjects of the instrument so that number changes.

I typically use XCMS (in R) to filter and do a retention time correction of all my samples, afterwards I normalize the signals.

With all that being said \x{2013} this is what I get after analysis\x{2026}

1. m/z
2. Retention time (from liquid chromatography), though not always, depends on the experiment that is being performed
3. Normalized signal for each m/z

- Goal: "In the end, I just want to know what RT and m/z differ across samples and which samples they differ against (A vs. C, A vs. N, etc.) so that I can work on identifying those peaks. Maybe an ANOVA would work since I have multiple groups\x{2026}."
- Statistical Inputs at clinics:
    - When acquiring the data from mass spectrometry, make sure your samples are randomized;
    - Assess the quality of the data: here are some possible ways to check: calculate CV within each group to check the "single to noise" ratio of your data; calculate the ICC to check the reliability/reproducibility of your measurements; apply clustering techniques to visually check if you can cluster all the control sample together;
    - For data with correlated structure (biological replicates, technique replicates), apply linear mixed effect model;
    - Clarify the goal: if the goal is to find individual "retention" time that can be a potential markers, univaraite analysis; if the goal is to do the prediction, may consider multivariable analysis, for example *lasso* based modeling techniques.

4 Dec 12

No clients

6 Nov 12

Ana De Lucas, Ophthalmology

- 96 well plates, EPO, proliferation assay; controls not placed randomly on plates
- Cell proliferation (more fluorescence)
- Increasing concentration of EPO - 4 levels
- Wild type vs. mutant
- 24h and 48h; one plate per time point
- Standard ANOVA is for comparing two or more experimental conditions applied to independent experimental units
- Assess quality of experiment by looking at variation across replicates; can use intraclass correlation coefficient (ICC) for this, or other measures
- Independent wells of cells for each time, concentration, genetic group
- 2x4x2 ANOVA, i.e., ANOVA with 3 factors (sets of conditions)
    - ANOVA will not handle concentration trend efficiently
    - Could assess Spearman rank correlation between concentration and cell growth
- Ordinary parametric ANOVA F-tests are not necessarily robust to strange data
- If time is allowed to interact with the other factors, this is equivalent to a separate analysis by 24h, 48h
- If do separate analyses by time, this is a 4x2 ANOVA if you did not use the ordering of concentrations
- Main question: is mutant resulting in more proliferation than wild type

- More specific: what is the difference in proliferation between wild type and mutant as a function of EPO concentration
    - 4 estimated differences with confidence intervals - do this in context of overall 4x2 ANOVA, i.e., 4 contrasts with simultaneous confidence limits to control overall confidence coverage of 0.95
    - One estimate of variability (residual variance) from a unified ANOVA
- May be problem with lower limit of detection
- See http://www.gene-quantification.org/malo-data-analysis-2006.pdf

30 Oct 2012

Dehui Mi, HTS, VICB

Consultants: Bill Dupont, Uche Sampson, Frank Harrell

I have no data set to send. What I need is someone clearing out my confusion in some very basic concepts, such as standard deviation, z score, B score, normal distribution, the difference of population and sample, etc. You can help me to clear out my confusion in these concepts so that I can talk about them correctly in my presentation.

- High throughput screening, 384 well plate
- How to identify a "hit" based on "very different", e.g., 2-3 SDs away from the mean
- SD is a measure of dispersion (variety); average absolute difference between two observations is a more intuitive concept (Gini's mean difference)
- SD is the square root of the variance and is on the original data scale
- n-1 is used in the denominator instead of n to give a penalty for having to estimate the center (mean) of the distribution; result is an unbiased (i.e., right on the long-term average) estimate of the population variance
    - The sum of squared differences is minimized when computed from the sample mean so it is optimistic when estimating the true variance
- Step 1 - primer screen - compare every single well to all the test wells on the plate
    - Plate has positive and negative controls on it in a checkerboard pattern; unknown test wells in center
    - Z' = 1 - 3*[SD(+) + SD(-)]/|mean pos - mean neg|
        - Unclear statistical justification for adding two SDs; in other contexts in statistics you add the variances then take the square root
    - See http://en.wikipedia.org/wiki/Z-factor and http://www.gene-quantification.org/malo-data-analysis-2006.pdf and http://jbx.sagepub.com/content/4/2/67
- Step 2 - confirmatory screen - compare wells to control
- Beware of methods that are tightly tied to the normal distribution
- Mean and SD are not robust statistics, i.e., a small number of "strange" values can distort their values

9 Oct 2012

Andy Link and Parimal Samir

To discuss our approach using semi-supervised learning (SVM) to validate peptide identification from mass spectrometry experiments. We are developing a novel target-decoy strategy. Decoy peptide hits are labelled as incorrect and target peptide hits are labeled as correct hits. The classifier is trained on a dataset containing both decoy and target hits. Finally, the trained classifier is used to validate the peptide identifications. We would like to discuss if our approach is statistically sound and get back your feedback.

Background

- most of yeast do not have much isoforms, mostly one gene, one protein

- isoforms - how exons are put together to make the final transcript

- yeast genes also not have much introns

- transcription factors can be far away from the gene it is modulating

- proteomics are most useful for figuring out protein interactions and modifications

SEQUEST

Goal: shotgun dataset, proteins are cut into peptides, to figure out what the peptides are, what the proteins are

Input: (1) precursor M/S ratio: MS scan - measures the mass to charge ratio, a measure of how well the peptide ionizes (2) MS/MS - a list of fragmented ions

Software: compare observed (1) and (2) values with theoretical values in the database

Proposed algorithm:

- hybrid of SVM and filtering

steps: (1) data cleaning: calculate centroids of targets and decoys, remove portions too close to decoys (2) SVM1: (3) SVM2: (4) filtering

Comments:

- "cleaning" training data may actually lose power for prediction on "uncleaned" test sets

- may want to define the utility function - a function of sensitivitiy, specificity and cost of misclassification

- ultimately only performance on test datasets matters, a classifier can perform superbly on tranining data but only ok on the test data

2 Oct 2012

Jennifer M. Giltnane, Dept. of Pathology, Microbiology, and Immunology

I have data structure/analysis question regarding how to best demonstrate change in endpoint marker (tumor proliferation before and after treatment) and compare it to an experimental finding (mutation). Data set is saved in ~/clinic/hddata.

Jun Dai, Epidemiology

25 Sep 2012

Qiuyun Fan, BME PhD student, affiliated with VUIIS and Education and Brain Research Lab

I'm working on one of the Kennedy Center sponsored projects. I have a dataset where I want to determine what factor(s) (targets 1-68) are important predictors of group/category difference. I first used PCA to reduce dimentionality, and then used linear regression to regress group membership onto the PCA scores. I then translated the beta in the PCA domain to original factors' domain to get an interpretable result. When I did linear regression in PCA domain, Matlab returned the confidence interval for each element in the beta. I'm wondering if I can translate the values of confidence interval to the original factors' domain as well.

- response: good reader vs. poor reader; better to use continuous score. Group is defined by a cut-off of continuous score, and subjects in the buffering zone are removed (this could cause bias).
- Predict brain function from brain structure
- 40 subjects (after 15 patients removed), 68 variables(neuroimaging pixel)
- 5 PCs explain 90% of variance
- Could explain the meaning of individual PC by examining the loadings.
- Examine the correlation between the variables and PCs
- SE of the coefficients of the original variables could be very big due to high correlation between the variables.
- Could group the original variables based on previous knowledge. Variable clustering based on correlation.

Doug Johnson, Hematology

- Does geno type affect the response to immunotherapy? Sample size justification
- Two variants of Ras
- Ras variant patients respond better than wild-type patients
- 20 Ras (9 responded), 31 wild-type (6 responded)
- Another endpoint: survival to death, or survival at one year
- What's the feasible maximum sample size attainable? At what time point will the response be observed? Confounders in retrospective study (Ras patients maybe more likely to get treatment)
- Obtain the information on the minimum detectable difference (in response rate) from previous literature
- Make power curve

18 Sep 2012

Consultants: Bill DuPont, W. Wu.

Joe Solus, Medicine

Seeking help in R01 grant development relating to miRNA arrays. Referred to A. Shintani, W. Wu.

Quinn Wells

GWAS SNP study for association with reduction in EF following chemotherapy.

- EF measurement modalities:
    - Nuclear - more precise
    - Echo - less precise, often censored recording (i.e., "greater than 55")

- Cumulative chemotherapy dose is related to EF drop. If EF drops due to toxicity, chemotherapy is stopped.
- Idea 1. Use a protocol to identify "cases" and "controls", based on the magnitude of drop in EF
- Idea 2. Subset on patients measured via nuclear technique, regress drop in EF onto SNP data.

28 Aug 2012

Susan Kroop, Rheumatology

Consultants: Bill Dupont, Dan Ayers, Frank Harrell, Ming Li, Bill Dupont

I am beginning an education project on housestaff rheumatology curriculum change. I am at the stage of creating my surveys for data collection and want to make sure the design of the surveys will give me the data that I want/need.

- Interested in residents' attitudes, knowledge, skills
- Survey before/after 1w rotation (n=45/y); compare one year with next year
- Main question is whether questions as composed will yield analyzable data
- Questions have not been previously validated
- Using REDCap analog scales mainly
- Advantages of brevity
- Think about biases caused by pre-post design (worst case: residents answer the way they think you want them to answer)
- Suggest talking to Irene Feuer
- Have multiple people look at draft questions

Thomas DiSalvo, Cardiovascular Medicine

Consultants: William Wu, Bill Dupont

Working on VICTR proposal looking at microarray gene expression in human right ventricles (RV) from explanted human myocardium. We are interested in comparing human RV gene expression (which has never before been studied, believe it or not) in 12 explanted heart failure hearts all of which have end-stage LV failure (hence the transplant...) obtained at time of transplantation and also in 4 control hearts (non-used potential donors). Our sample sizes are "fixed" by the number of prepared samples - 12 cases, 4 controls. All 12 cases have end-stage LV HF - hence the transplant. We've completed a proteomics study in the same hearts, so are limited to this number of samples to do correlative genomics-proteomics.

In this pilot study, we'd like to look at the microarray data with 4 possible aims:

1. Does RV gene expression differ from LV gene expression in normal explanted hearts ? 4 control RVs vs. 4 control LVs
2. Does RV gene expression differ from LV gene expression in end-stage LV HF explanted hearts? 12 pooled case RVs vs. 12 pooled case LVs
3. Does RV gene expression differ by etiology in end-stage LV HF explanted hearts? 6 ischemic RVs vs. 6 non-ischemic RVs
4. Does RV gene expression differ by RVEF in end-stage LF HF explanted hearts? 4 normal RVEF cases vs. 8 abnormal RVEF cases

Our analysis will involve both IPA and KEGG for molecular systems/pathways, but also for individual genes of interest, ANOVA with Benjamini/Hochberg correction and "q value" is our tentative plan (at least as is planned now). Given the small samples size and "shifting" samples sizes for and possible analyses, want to ensure that before the VICTR application is reviewed, it's reasonably argued and feasible. If you could also suggest a biostatistician with particular expertise in analysis of "smallish" micro-array pilots, would appreciate it as well. Using Affymetrics whole transcriptome array. Usually 250-500 genes show differential regulation.

- Adequacy of sample size will depend on cross-subject variance of log gene expression among other things
- Pathway analysis is probably necessary, vs. individual gene screening
- RV and LV intrinsically paired; n=15 or 29
- Controls are varied; n=4 probably futile; large false negative rate
- VICTR applications can request specialty biostatistics support (here, genomic analysis) or any amount; More than $2000 (20 hours) must be matched 1:1 with funds from home Division
- Estimate of time required: 30 hours ($3000 voucher request)

Tricia Thornton-Wells, Jennifer Vega - MPB, Neurosci

Consultants: Hakmook Kang, Bill Dupont, Frank Harrell, Dan Ayers

- African Americans, majority family hx Alzheimer's disease
- Neuropsych measures of cognition to be correlated with functional connectivity measures (within-group design)
- Also interested in between-group design
- Initial VICTR application used standardized effect size in power calculations
- Is it possible to use an effect size in real units? Or base calculation on precision of parameter estimates?
- Regression framework: Y = cognition X = connectivity measures; need small # connectivities of interest (may limit it to one)
- Can compute sample size needed to yield a small margin of error in estimating any one correlation coefficient (n will be at least 100)
    - PS software will compute power for a slope

- Still not clear how to state the final result in biologic or patient-meaningful units

Razmia Alawi, Nursing

Consultant: Frank Harrell

- High turnover of nurses on gen surg floor
- Who to interview (prob. floor managers) and when to interview them
- Think about 2 stages: one qualitative phase to get universe of answers from some of the mangers, one to ask all managers to rate the applicability of these answers to what they've seen
- Suggest talking to Ken Walston, Warren Lambert, Irene Feuer

14 Aug 2012

Elizabeth Pearce, ENT resident, Dept. of Otolaryngology; PI David Francis

Conducting a prospective study for which we need Biostats help both with formatting the statistical design, and for a price quote for Biostats consult for eventual data analysis. it is a prospective study, with testing once before and once after surgery for vocal fold immobility, to measure multiple parameters, the most important of which is respiratory rate during a treadmill test. We spoke with Dr. Christopher Slaughter who recommended repeated measures statistics, which we are unfamiliar with doing on our own. However, we are comfortable doing stats for the other measures in our study.

Study Design: This is a prospective, feasibility study of patients with Unilateral Vocal Fold Immobility (UVFI) being treated with injection laryngoplasty procedure. We hope to have 10 patients complete study, enroll 20-25 (need to do power calculation). Patients will undergo testing once pre-operatively, then once, 3-weeks post-operatively. UVFI can worsen voice, breathing, and swallowing due to an insufficient glottis (voice box). The injection laryngoplasty is a standard of care procedure designed to improve glottic closure. Our primary outcome measure is Respiratory Rate (RR, breaths/minute) during an exercise tolerance test. Patients will serve as their own control, comparing pre- and post- treatment results.

Aim: Comprehensively evaluate patient respiratory function pre- and post-injection medialization.

- Our primary outcome measure is change in Respiratory Rate (RR) during the **exercise tolerance test** (see description below) between pre- and post-treatment time points.
- RR: In healthy subjects, RR range from 3 breaths per min (br/min) to 28 br/min. The mean rate was

1. 49 br/min with standard deviation of 4.36 br/min. (Addison PS, Watson JN, Mestek ML, Mecca RS.An algorithm for pulse oximetry derived respiratory rate (RR(oxi)): a healthy volunteer study. J Clin Monit Comput.
2. Feb;26(1):45-51. Epub 2012 Jan 10.)
3. We will also measure other outcomes, listed in Table 2, yet RR is our primary outcome. The other outcomes are dichotomous measures and we are comfortable doing these less complicated stats.

Hypothesis: Pulmonary function, measured as Respiratory Rate during exercise test, will decrease (improve) after injection because it enables patients to efficiently use the Valsalva maneuver to control intrathoracic pressure and breathing.

Exercise Test :

Patients will walk on a treadmill using the Bruce Protocol (Table 1), with incremental increases in speed and incline every three minutes. Data collected: Our primary outcome is to monitor is respiratory rate (breaths/minute). This will be measured once before the test, then at the last 20 seconds of each 3 minute stage (7 stages total if patient completes entire protocol), then lastly, once at the end of the treadmill test for 9 total time points. We will also measure: Heart rate, Blood prssure, METs, Oxygen saturation, total time on the treadmill, and highest Bruce Level achieved. Currently, there is no normative data in the literature about RR for the treadmill test, or patients with UVFI.

Table 1: Bruce Protocol Exercise Treadmill Testing

Stage	Minutes	% grade	km/h	MPH	METS
1	3	10	2.7	1.7	4
2	6	12	4.0	2.5	6.6
3	9	14	5.4	3.4	9.1
4	12	16	6.7	4.2	12.9
5	15	18	8.0	5.0	15.0
6	18	20	8.8	5.5	16.9
7	21	22	9.6	6.0	19.1

Table 2: List of Tests Performed Once Pre- and Once Post-Treatment

The table below outlines tests performed within each of the three major parameters: Swallowing, Voice, and Respiratory. (S) = Standard of Care, (R) = Research (**in bold**).

Swallowing Parameters	Voice Parameters	Respiratory Parameters
Dysphagia Handicap Index (DHI) (S)	Voice Handicap Index (VHI) (S)	Pulmonary Function Tests (PFTs) (S)
	Sustained Phonation (S)	***SF-36 Quality of Life Survey (R)***
	Stroboscopy (S)	***Chronic Respiratory Questionnaire (R)***
		10-point Dyspnea Scale (R)

		Exercise Treadmill Testing (R)
		Post-Exercise Survey (R)

- Pilot study based on convenient sample size (N=15), wanted to calculate power
- Have repeated measures data. Simplify the power calculation based on paired t-test. Can use PS software.
- Need to control the ability (at baseline, how many blocks they can walk), which associates with how far they can go during the study.
- Multivariable model, 1:15 rule (need 15 subjects for each covariates included in the model)
- Will apply VICTR grant, estimate 40 hours of work (~$4000).

24 July 2012

Martin Schmidt, Psychiatry/VKC, Adam Anderson, BME

- Mass spec with mass/charge ratios, by multiple regions or voxels
- Has functional data analysis of protein spectra been extended to this setting?
- Discussed advantages of unified modeling that respects spatial structure
- Wavelets are worth thinking about

10 July 2012

Ryan Delahanty, Epidemiology

- `aregImpute question` : https://dl.dropbox.com/u/9445847/CNV.Imputation.R

Robert Turer, Trent Rosenbloom, DBMI

- ICD9 vs ICD10; important to map between, to e.g. compare on old data
- ICD10 more specific, includes laterality, more procedures
- GEMs: general equivalence mappings; include 1:1, 1:many, combinations
- No prospective research examining coder-generated 9, 10 codes vs. GEMs: validation of GEMs
- May want in the future to pick at random two coders for each chart, to be able to study inter-observer disagreement and see if this amount of disagreement is in the same range as coder vs. GEMs
- Current data not allow this
- Choose 100 cases, hopefully diverse
- weighted Kappa

3 July 2012

Genie Hinz, DBMI

Evaluation of four different risk scores. To evaluate the benefits and limitations of the survey version as compared to the electronic version.

- missing information on ADL data from 44 patients.
- Create 2by2 table for survey and electronic version separately using review chart as a gold standard

24 April 12

Adeline Dozois, Brian Cash, Gadini Delisca, Alejandro Perez, Pooyan Rohani, Emily Zern

Research question: For patients presenting to the Accident and Emergency (A&E) Department of Georgetown Public Hospital Corporation (GPHC) with Staphylococcus aureus infections, is there an elevated level of IgG antibodies against two virulence S. aureus virulence factors (lukA/B and \x{03b1}-hemolysin) following a four-week convalescence period relative to the IgG level during acute infection?

Study design: Blood samples will be collected from 150 patients visiting the A & E of GPHC over a four-week period. These subjects will be asked to return in two to four weeks for a repeat blood draw. Blood samples will be spun at the GPHC clinical chemistry laboratory and divided into two aliquots. One aliquot will be stored at the GPHC, and one will be sent in batches to Vanderbilt University. Paired serum samples will be analyzed by ELISA to determine total IgG concentration, as well as IgA concentration measured against two specific S. aureus virulence factors.

Sample Size : 150

- Guyana main public hospital; 60% of African descent, 40% Indian -southeast asian
- Prevalence in staph aureous infections of resistance
- Immunological analysis from blood samples; 4w later additional blood sample
- ELISA IgG non-spec plus 2 specific vir. factors; describe antibody response; are antibodies still there 4w later?
- Planned paired t-test for acute vs. recovery rise
- Want to compare with VU population
- Bland-Altman plot: Y=post-pre; X=post+pre; desire: random scatter that's flat
- Alternate: log (post/pre) vs. geometric mean ... log(pre) + log(post)
- Or: square root or cube root

- Check for difference being on the right scale, i.e. transformed variables properly
- Decide on transformation from 4 plots; transform -> take differences -> paired t-test or Wilcoxon signed-rank test (latter somewhat preferred)
- Population differences: double difference: compute f(post)-f(pre) in two populations; compare using Wilcoxon-Mann-Whitney rank sum (unpaired) two-sample test
    - The latter test is independent of the transformation
- Sample size: need raw data, quartiles, or SD
- Tertiary analyses: regress age, weight, sex, ... on f(post)-f(pre) (multiple regression)
- Or: regress age, weight, sex, country ... on f(post)-f(pre)

17 April 12

Alexandra May, Trisha Pasricha, Ian McGuinness, Richard Samade, David Amsalem, Zain Gowani

-

21 Feb 12

Genie Hinz, Biomedical Informatics

- Look at outpatient populations, 4 year mortality rate on a sample around 3000 (identified pts with specific physicians)
- Suggest keep it as continuous as to divide pts into low, intermediate, and high risk
- Used age, gender, etc. to define the pts as low, intermediate, and high risk, if the primary interest is the risk, then no need to adjust for other risk factors
- One of the biggest confounders is activity daily living, with 0/1 reading, while in another study ADL has scores from 0-7; Why got similar ROC? Might because all other variables already explain the variance,

14 Feb 12

Matthew Duvernay, Pharcology (mentor: Heidi Hamm)

- VICTR application regarding sample size and statistical analysis plan, needs more detail
- Paired samples: PAR1 stimulated vs PAR4 stimulated, compare protein expression
- Need to know how many subjects needed to detect statistical significance
- Outcome: mass spectrometry intensity
- Suggest take log2 transformation and use paired-t test to calculate the required sample size

- Analysis: Use Wilcoxon signed rank test to compare within subject, which does not assume normal assumption
- Get at least 3 samples from pilot data to get an estimate of standard deviation

17Jan12

Special Clinic: PREDICT

- Josh Peterson, Kevin Johnson, Marc Beller, Ioana Danciu, Jennifer Mitchell, DBMI
- Biostatistician discussants: Bill Dupont, Frank Harrell, Cindy Chen, Jonathan Schildcrout
- Understanding phamacogenomic effects
- Preemptive genotyping, starting with clopidogrel for patients getting a stent
- Topic today is evaluation of the program
- What is response of end users when they get the new information?
- There is a variety of both efficacy endpoints and safety endpoints to consider
- Related to clop. metabolism there are 20-25% heterozygotes and 2% homoz.
- What simulated decision analyses will help?
    - Any simulation will have uncertainty that is limited by the literature's margin of error in estimating differential treatment effect, no matter how many "patients" are simulated
- Cost of genotyping vs. cost of using another drug
- Information base for differential treatment benefit (strong pharmacogenomic hypothesis) is now confusing for CYP2C19
- Challenges of retrospective evaluation (e.g., changes in concomitant therapy) vs. prospective (cat is out of the bag with respect to FDA genotyping recommendations)
- How would one design a good randomized clinical trial? Genotype everyone, mask random half
    - parallel group
    - matched cohort
    - retrospective case-control study - genotype upon MACE (major cardiovascular event) plus controls; may have to avoid using VU
    - observed estimate of efficacy is a function of:
        - relative efficacy and safety of clopidogrel vs. prasugrel
        - genotype-differential benefit of prasugrel
        - average baseline risk of major cardiovascular events (if study is prospective)
        - how the different endpoints are weighted (esp. bleeding)
    - C4PG or Case4PG web site deals with some of these factors
    - some of literature only provides (fuzzy) estimate of differential benefit of clopidogrel over placebo
        - there is one good paper in the literature for clop. vs. pras. by genotype
    - if prospective, completeness of follow-up is imperative

20Dec11

Raafia Muhammad, Cardiovascular Medicine and Lan Jiang, CHGR

- Went over BioVU application and developed a plan for validating 100 SNPs found in a discovery cohort
- Recommended plotting log odds ratios from discovery cohort vs. log odds ratio from validation cohort
- Need to clarify whether to adjust all these for age, sex, etc.

Rachel Lippert, MPB

- Analysis of race effect on SNPs, looking at low BMI
- SNP has been identified in the 1000 genome database
- Make sure any selection biases in cases and controls are similar by design
- Suggested 1-1 case-control ratio, using low and high BMI

25 Oct 11

Katie Hutchinson, Graduate Student, Cancer Biology

- Study of genetic mutation of cancer. One tumor sample, one normal sample. Want to compare the distribution between two samples.
- Suggest using graphic display of the data of the two samples
- Circos: data visualisation for the genomic data

2 Aug 11

Ben Shoemaker, Maureen Farrell, Cardiology

- Went through Cox model analyses with updated data - significantly longer follow-up
- May be a good idea to get more patients; current data can be used to plan how many, depending on follow-up schedule

26 July 11

Ben Shoemaker, Maureen Farrell, Cardiology

- Implantable cardiac defibrillator (ICD) before ventricular tachycardia/fibrillation (VT/VF) (primary prevention)

- Refining criteria for selection of patients for ICD
- Genetic marker
- DISCERN : multi-center trial, VU was a center; SNP 4q22 p=1e-8 related to axon development
- Validation cohort from another study; primary prevention subset; completely independent of discovery cohort (no sample contamination)
- Need to extend follow-up of all patients, especially to handle the problem with one patient having an event at 745 days when others were not followed but 730 days
- Went through Cox model analyses and issues with covariate adjustment
- See if one missing LVEF can be filled in

5 July 11

Yosaf Zeyed, Pulmonary, Allergy, Critical Care Med, Dept. of Med.; MSCI

- Applying for VICTR funds - gene micro RNA in acute lung injury
- Aim 1: 5 groups (ALI w/sepsis; ALI with trauma, sepsis without ALI, trauma without ALI, control - from ICU without any of those)
    - 6 pts/group
    - Outcome has already happened; serum samples pooled because of cost
    - If do arrays individually by patient, cost is about $15,000, possibly feasible with VICTR funds
    - Up to 1000 candidate features/markers; this is the dimensionality of the problem
- Aim 2: screen for specific micro RNA
- Discussed problems with high false negative rate
- Could screen markers based on uniformity of expression in 6 patients within the group, then look to see if different groups have differently homogeneously expressed markers
- Note that the more funds requested from VICTR the more the mentor has to be involved in the application

28 June 11

Jamie Ausborn, cancer biology

- Four cell lines, three replicates of each, microarray

24 May 11

Sarah E. Williams, Ronina Libeth, Pediatric Clinical Research UURP

- Bill Dupont went through the use of PS software, discussing relationship between power, detectable difference, etc.
- Need an estimate of the standard deviation across patients

3 May 11

Phil Lammers (Department of Medicine, Hematology, Oncology)

8 Feb 11

Satish Raj (Clin Pharm) and Kirsten Haman (Psychiatry) --- "Origins of Cognitive Dysfunction in Postural Tachycardia Syndrome (POTS): A Pilot Study"

- Biostatisticians in attendance: Bill Dupont, Pencheng Lu, William Wu, Lily Wang, Theresa Scott
- VICTR submitted protocol that went through pre pre-review (Frank was reviewer); asked to attend a Biostat Clinic.
- Did not have list of specifics that needed to be addressed, but discussed general improvements, which included:
    1. Modify plans for "matching":
        - Instead of "gross matching on intelligence", exclude those subjects/patients who score below a population level intelligence from the analysis.
        - So, will only be matching on age.
            - Dr. Dupont mentioned using "frequency" matching instead of individual matching.
    2. Providing more detail of where control *subjects*will be recruited from.
        - NOTE: protocol refers to cases as "patients" and controls as "subjects".
    3. Provide more clinical reasoning for using the z-scores in the analysis --- common analysis? how z-score calculated? how reproducible?
    4. Provide more clinical reasoning for using the "2SD" cutoff --- is this cutoff calculated from the sample or is it more from a population/clinical standpoint.
    5. Possibly using regression analysis to analyze those with normal or abnormal (based on 2SD cutoff above) composite score between cases and controls --- instead of comparing using just a Chi-Square.
        - Adjusting case/control effect for continuous composite score.

4 Jan 11

Ken Monahan, Cardiovascular Medicine: Sleep Apnea and Atrial Fibrillation

- VICTR grant nearing end
- Need sample size for larger study
- Genetic contribution to propensity for Afib given sleep apnea
- Using BioVU; 1500 with blood banked, EKG, sleep study
- Limitations of sample: many who didn't get sleep test have sleep apnea; many who get sleep test will have SA (enriched sample)
    - Think about target population and how can you approximate the estimates that would be obtained from the target pop.
- Model for probability of (binary) Afib | genetic + clinical factors
    - Model with clinical factors; add sleep apnea factors; add genetic factors
- Limiting aspects of sample size: number of Afibs in the 1500; allele frequencies
- Can take a prediction approach
- Candidate predictors: 10 clinical/demographic, 5 sleep study, 30 SNPs -> 15*(10+5+30) Afibs = 675
    - 15:1 rule of 15 events per candidate regression coefficient comes from simulation studies of how many outcome events are needed per candidate variable in order to fit a model that is as good as it apparently is
- This is assuming that all SNPs are given equal prior importance
- If SNPs are restricted to have 10 effective d.f., would need 15*(10+5+10) = 375 Afibs
- To answer a question about genotype x apnea interaction would probably require many more subjects
- Alternative: something like group lasso where one finds the optimum cross-validating shrinkage of the 3 types of variables (3 shrinkage coefficients)
    - SNPs shrinkage factor of infinity -> genetic information can be ignored
- Crude estimate of biostat cost $8000

14 Dec 10

Matt Landman

VICTR proposal - mutational analysis of hereditary pancreatitis pedigrees

23 Nov 10

Yan Ru

VICTR proposal 1288

genetic profiles using Affymetrix cDNA arrays, using normal, MMR, IMR groups

5 samples for each grp, id differentially expressed genes, will verify with PCR

about 28000 genes will be examined

suggest:

- when comparing groups, need to account for within group dispersion. Can use T statistics rather than fold changes, which only account for changes in means
- need to evaluate false discovery rate. 10% cut off.
- need to increase sample size to at least 10 to 15 per group. (Original proposal is 5 per group for three treatment groups.)
- One strategy would be to use only two groups, normal vs. MMR, and increase the number in each group.
- factors to consider: will be publishing for this pilot data? feasibility assessment needed? effect sizes expected to detect?

2Nov10

David

PCA analysis - the first principal component score is the linear combination that explains largest amount of variation in gene expressions among all linear combinations so

Amanda

Imaging problem

- need to do a linear regression with intensity as outcome, group as independent variable, and date as covariate variable

Dale Tylor

- retrospective cohort study - VICTR voucher request is being made

26Oct10

Andrew Link, working with Kathy Edwards, Pediatrics Infectious Disease

- Reponse to vaccination: novel biomarkers
- Goal is to have a test for whether the vaccination protects a subject
    - Without waiting for 7 days, for example if antibodies are tested
- Proof of concept by comparing vaccinated and non-vaccinated subjects
- Query immune system after vaccination
- T-cells, memory cells expensive to measure
- Blood sample 1, 3, 7d after vaccination; analyze neutrophils, B-cells, T-cells, ...
- Profiled for changes in transcriptome (RNASeq) and proteome (mass spec)
- Cytokine responses
- Naiive patients used in first round, to try to deal with already-protected subjects
- Could require pre-vaccination titer to be below a threshold
- May be beneficial to study a vaccine that is 50% effective - wouldn't need controls
- One way to get a handle on the multiplicity/lack of validation issues is to generate a random matrix of the same dimension as the real data, and to do all "real" analyses in parallel on the random data; if one obtains an ROC area of 0.85 on the real data and 0.55 on the fake data, then we have more confidence in the results. However if the ROC area achieved with the fake data is also 0.85 we have to worry.

5Oct10

Chantel Sloan, Division of Allergy, Pulmonary, Critical Care, Dept. of Medicine

- Air pollution data over time, spatial modeling, spatial interpolation, modeling different land uses
- Consider data reduction methods such as missing data PCA
- Can have Chris Fonnesbeck (ecological statistician) attend a future clinic
- Be sure to factor in uncertainty in alphas for all later steps

Adrienne Dula, Radiology

- Came to last Thursday clinic
- New imaging technique; interested in sensitivity and specificity
- Gold standard is as used in clinic
- Goal: distinguish tissue types, e.g. white vs. gray matter
- For a career development grant
- Current plan - overlay segmented image on new method
- Can you register using histology?
- Can one use ratios of contrasts?
- Discrepencies between edges, dealing with one method not finding an edge
    - Might use rank measures; not detected = worst rank

Xue Yang and Bennett Landman, Electrical Engineering & Computer Science

- Related question; how is one image related to another
- Comparison of two random variables; Y, random Xs, fixed Zs (e.g., demographics); Y | X, Z
- Nonparametric voxel by voxel regression but worried about random Xs
- Likelihood P(Y, X | Z) = P(Y | X, Z) P(X | Z)
- Could fit a time series model followed by a group analysis both using SPM
- May start with (Y - X) | Z; if multiple images (not just two), can analyze all possible pairs (or k-1 pairs for k images) and use the cluster bootstrap to get confidence intervals taking into account what information overlaps

14Sep10

Sarika Peters, Dept. of Pediatrics and VKC (VICTR pre-review clinic)

- Consultants: Frank Harrell, Warren Lambert, Lily Wang, Cindy Chen
- Preliminary data - imaging study - children with Angelman's syndrome - 14 patients, 13 controls - ages 8-17
- To be used for a Feb11 R01 submission
- To use same imaging techniques, eye tracking, event potentials
- Need to demonstrate feasibility of conducting study at VU with a study team in place
- Target 10 patients to get 6 (2 + 2 + 2); issue of need for sedation during imaging (50 min.); seeing about one pt/wk
- 3 syndromes - different genetics, similar phenotypes, but involve different brain areas
- Goal is feasibility, not group differences
- Did analyses using 2 independent imaging interpreters, estimated inter-rater reliability
- Brain regions of interest pre-defined
- Difficult to deal with medications patients are on (e.g., anti-seizure)
    - In ultimate analyses might think about correlating key variables of interest with time since last dose; can also look at which medications were used
- For R01 imaging will be at baseline only
- Asking VICTR to pay for scanning, sedation, ERP, work towards power/sample size for R01 based on first preliminary data (N=14 + 6 patients)
    - VICTR proposal doesn't need to ask for much biostat time
- Resources: VICTR, VKC Statistics and Methodology Core , Dept. of Pediatrics biostat resource (senior statistician: Ben Saville)

Ileko Mugalla VIGH question from 13Sep10 clinic

- Who are resources in qualitative research? Sabina Gesille in Gen Peds, prople working with Len Bickman in Peabody; Warren can be contacted to help fine more collaborators

31Aug10

- Ehab Kasasbeh, Cardiovascular Medicine - see FridayClinicNotes(consultant: Bill Dupont)
    - Demonstrated PS software for computing power to detect specific odds ratios in the two-sample binomial problem
- Borden Lacy, Stacey Seeback, Microbiology and Immunology (consultant: Fei Ye)
    - Question about displaying standard deviations
    - Potential problem with analyzing on the percent or ratio scale
        - May be able to compute confidence intervals on the log scale, anti-log to get fold-change confidence intervals
    - Be sure to display raw data
    - Issues of within- and between-plate normalizations
    - Not a good idea to divide by mock treatment - display groups, not one group "normalized" for the other
    - Need a full model that will take into account the variability in the mock treatment group
    - In the first phase of the SI genome project want to pick genes from 21,000 genes
        - Depending on biologic and technical variability, there may be a significant problem with false negatives
        - Pilot study being used to refine technique, reduce variability
- Bennett Landman and Baxter Rogers, Radiology and VUIIS (consultants: Frank Harrell, Bill Dupont; Lei Xu is out of town)
    - Neuroimaging studies are very expensive
    - EMR can be used for retrospective analysis
    - Does a treatment outcome correlate with something that could have been discovered by imaging
        - E.g. subregion size, shape, amount of fluid surrounding the brain
    - Also interested in creating atlases showing normal biologic variability; anatomical and resting state; PET as static capture of function
    - It may be possible to contribute images to BioVU
    - Images come with a set of parameters describing the data acquisition/hardware
    - Lossless compressed image data will be maintained as a separate database linkable with the SD
    - VICTR Design Studio suggestions
- Babar Parvez, Cardiovascular Medicine (consultants: Bill Dupont, Pengcheng Lu)
    - 399 atrial fib patients who were rhythm-controlled
    - Literature identified 3 genes (4 SNPs) related to increases risk of afib (through ion channels etc.)
    - Do these have to do with altering the effect of meds (rate vs. rhythm control meds, etc.)
    - Polymorphisms thought to mainly effect rhythm control
    - Have the results for the 4 SNPs
    - Y = success/unsucessful rhythm control
    - Traditional to combine various genotypes; but what to assume for heterozygous?
    - Two-parameter logistic model will allow the entire spectrum from recessive to dominant

    - Could also do an allele-based analysis (related to minor allele frequency); a 2x2 table analysis based on alleles is equivalent to a person-based binary logistic model that places heterozygous exactly halfway between aa and AA
    - Spoke about potentially increasing the study's power by turning the binary response variable into a continuous or ordinal variable

17Aug10

Jongchan Kim, Dept. of Pathology

- Known loss of function mutations
- # copies > 2 vs. 2; # copies not available in data (which are from the literature)
- Would have been far more powerful had we had # copies or level of protein expression
- Need to know when metastasis was determined (baseline vs. follow-up)
- Survival time from date of diagnosis (all cause death); need to confirm is in days (but how did fractions of a day come about?)
- Need time to last known alive if not died, plus death/censoring indicator
- PI:Sarki Abdulkadir - have him come to a future clinic

Queen Henry-Okafor - Cardiovascular Medicine

- Biomarkers on acute decompensated heart failure on patients presenting to ED
    - gout or gout-related arthritis patients excluded (would have been on the drug)
- Uric acid of interest; does lowering the level reduce the need for re-visiting ED
- E.g. 7 mg/dl to 6.5 mg/dl in 3 months by putting on drug
- Starting sample: all qualifying patients who had a baseline and a 3 month visit
- Check association of uric acid change with later ED visit events
- Two ways to state the hypothesis:
    - Does the medication help (would need some patients randomized to not get the medication)
    - Whatever the medication does, does the resulting change in uric acid explaining a reduction in future events
        - Can't say that the drug did it (as opposed to a temporal effect/natural history of HF)
    - Doing the second study could lead to funding to do the second
- Can the level of HF be measured at the initial and 3m visits? Can correlate 2 uric acids and 2 HF measures (e.g., change vs. change)
- For sample size estimation for a randomized trial (drug vs. placebo, Y=HF symptoms e.g. ED re-visits, other HF assessments)
- Need to refine the outcome scale: how many levels, estimate fraction of subjects in each level
    - The more continuous the HF outcome scale the lower the sample size

- If only use the time until the first ED visit after 3m, need to estimate the incidence of ED visits over time

22Jun10

Jon Forbes, Neurosurgery

- Electrical activity of slices of rat brains; stroke model; stop oxygen perfusate
- After signals stop, restart perfusate, electrical signals back in 6m. Elect. stimulation at reperfusion -> faster return
- 2 groups: stopped, wait for signal cessation, observe for spontaneous recovery; stopped, signals lost, restore perfusate, e-stim; also record time to signal loss
- Y: Time to recovery of signal; no animal failed to have signal return (because of perfusate)
- n=30 rats (15 each group)
- Box plots with raw data are a good idea; decided on just dots

tco.lost

pdf('/tmp/p.pdf', width=5, height=3.5) stripplot(group ~ rec, jitter.data=TRUE, factor=.5, xlab='Time to Recovery of Baseline Electrical Activity (min.)', panel=function(...) { panel.bwplot(...); panel.stripplot(...) }) dev.off()

25May10

Erik Boczko, DBMI

- Measuring information content
- Two fluid samples from same patient; each sample used for same series of correlated tests
- Goal is to see if a non-invasive test is accurate. The genetic analysis can rank order concentrations better than estimate absolute values
- Testing 7 pathogens, and a universal bacteria test. Culture takes too long; Coulter counter cannot differentiate staff aureous etc.
- Need to show equivalence between non-invasive and invasive assessments
- Interesting to look at invasive test ordering tendencies; if invasive test ordering is deterministic, study could still be useful for differential diagnosis of organisms; otherwise there may be a problem with invasive test non-orders.
- His major concern at this point is assessing false negative results from his non-invasive test. His approach appears to be reasonable. As he is somewhat concerned about confidentiality, it is probably not appropriate to provide more details here.

18May10

Amanda Solis, Dept. of Microbiology & Immunology

- Confocal imaging of HIV particles in cell culture media; immunofluorescent staining; intensity reading on red = antibody
- How well is antibody binding to certain types of virus
- Comparisons of interest: For one type of antibody is there a difference in binding between two viruses, and for another type of antibody is there no difference in binding? But interested in answering separate question for each antibody.
- 3000 intensity readings per sample; geometrically random but can't say that antibodies attacking one virus are not the same antibodies counted for another virus
- Starting from the same virus culture have repeated the entire experiment 3-4 times
- How do two samples differ in intensity readings? Comparison of virus types.
- Could pool all data but distinguish the replicates by having a 1-4 "replicate variable; allows checking for replicate effort
- Recommend having the instrument output intensities in a 10x10 grid; argue that there is very little overlap in particles between adjacent squares, so effective sample size is 100 (times 3-4)
- Test for difference in viruses:
    - t-test if take sample size to be 3-4
    - Wilcoxon-Mann-Whitney two-sample rank-sum test (short name: Wilcoxon test) if use 10x10 grid (Wilcoxon test can't give small P-values unless n is a bit larger than 7 in both groups combined)
    - No binning of data is needed
    - Could use a regression model to test whether there are any geometry effects (e.g., add as a variable to the model the distance from the center, distance from the closest edge, etc.)
- Gain more information about distributions by plotting histograms or cumulative distribution functions of intensities over individual viruses

2Mar10

Discussion of Biomarker Discovery and Validation Strategies

- Should the validation measure be the same statistic as the one used to find the markers and adjust for covariates?
- Example: LR chi-squared for finding "winning" markers, use likelihood measure for measuring predictive discrimination
    - Binary logistic regression: logarithmic scoring rule (equiv. to -2 log likelihood minus a constant from a model with no covariates)
        - Note: quadratic scoring rule is Brier score
- General case: Nagelkerke R^2 or "Adequacy index" (measure of relative information content)

- Development of predictive model uses dose-response relationships
- Validation should likewise use dose-response relationships
- Example: develop risk score S using Cox model; using linear combinations
- Relate S to survival in a validation sample: need to show that mid S values correspond to patient prognoses in the mid-range
- Dichotomization of S at a point c implicitly assumes a discontinuous relationship that is flat on either side
    - More importantly it assumes that c is a true inflection point or otherwise is a risk threshold; if c is a sample quantile or mean, it is derived without reference to risk at all
- If S was derived from a continuous function need to validate it continuously and also show that middle values correspond to middle outcomes
- Calibration (reliability) curves are also of interest
- In many cases, it is of interest to partition S into S(clinical) + S(biomarkers) and show the variation in outcome if you vary S(biomarkers) and fix S(clinical)
- Continuous approach removes temptation to find alternative cut points c if initial validation is disappointing
- Can also consider measures (e.g., slope shrinkage) of how different a re-calibrated curve is from the original calibration curve from the training sample * There are measures of pure discrimination ability (C-index = ROC area; discrimination index related to R^2)
- Simple measures can also be used, e.g., histogram of predicted risk, scatterplot of predicted risk based on clinical variables alone vs. predicted risk based on clinical variables + biomarkers

15Dec09

Buddy Creech, Pediatrics Infectious Diseases

- Hemoglobin variation and staph aureous/disease severity
- Clinical outcome is invasive vs. non-invasive (in bone) disease
- Interested in applying for BioVU usage
- 12-14 hemoglobin residues of interest
- Plan on a discovery phase and a validation phase
- Would get more power with a continuous or ordinal outcome variable

24Nov09 1Dec09 15Dec09

Merida Grant, Psychology, Arts & Sciences

- Applied for VICTR funds; was asked for clarification re: data analysis plan
- Discriminative conditioning in unipolar depression
- 3 groups: depressed with and without hx of childhood trauma, age+gender-matched controls

- No fMRI data collected yet
- Frequently extract beta weights from Brain Voyager and feed them into SPSS
- 3 main brain regions of interest; want whole brain (voxel-wise) analysis and region-wise (should use one of these 2 for sample size analysis)
- Need to specify the difference desired to be detected; difficult on unitless fMRI data
- fMRI data: subtract baseline from active task
    - Need to do Bland-Altman-like plots to verify that the data have been properly transformed before taking differences, i.e., differences are independent of averages
    - Make sure that any normalization method used has tested the assumption that division is the correction normalizing operation and that log ratio should not be used instead
- Find literature to get estimates of standard deviations; will look for similarities (stimuli, imaging technology)
    - May simplify to a two-group problem
    - May have to assume that SD in the control group is the same as in the other groups
- If it makes sense to log values, or otherwise use a relative (e.g., fold change) measure of effect, then the power calculations are simpler
    - This assumes the literature provides SDs on the appropriate scale (e.g., log of raw values)
- Otherwise can approximate an absolute mean difference to detect by multiplying the mean by for example 0.2 (20% change)
- Watch out for variance in clinical samples tending to be much larger than variances in controls
- 1Dec09: Lei has a copy of a pertinent article comparing two groups (placebo, corticosteroid); use discriminative conditioning
    - Compared contrast CS+ - CS- across the two groups
    - Paper makes the common mistake of showing a dynamite plot with error bars not corresponding to the experimental design
    - Need SD of within-subject CS+ - CS- differences across subjects
    - Otherwise need to estimate the correlation coefficient for CS+ vs. CS-
    - But also provided the maximum t statistic over the 32 voxels within a cluster; t-stat could be large just because the SD was underestimated in a voxel
    - Better to use region of interest average as the basis for designing the new study, but this is not available from this paper
    - Simple approach: solve for SD from the maximum t and use Bonferroni multiplicity adjustment when computing power; hope that Bonferroni is conservative enough to account for the downward bias in SD; i.e. use alpha=.05/32 = 0.0015625
    - Need the one voxel mean difference one would not want to miss
    - t = 3.01 = [(1.4 - -.2) - (.2 - .4)] / [sigma * sqrt(1/20 + 1/28)] = 6.148 / sigma; sigma = 2.043
    - **NOTE**: This may pertain to ROI
    - What is required is the biological difference one would not want to miss, on the same scale as the 2.043
    - For example if the delta to detect is 0.5, and alpha=0.05 power=.8, n per group must be 263
    - Used PS software available from PowerSampleSize 15Dec09

17Nov09

Tracy McGregor, Pediatric Genetics

- Planning sample size for validation SNP study in scoliosis
- 17 SNPs survived discovery phase
- Power vs. precision (of odds ratios) approach; case-control study

1Sep09

See Hardesty 18Aug09

George Jules, MMC

- Question about microarray analysis and GO

25Aug09

George Jules(Heart), Supervisor: Darryl Hood

- Look at the raw data by logging in VMSR Vanderbilt, think about changing platform to Affy
- Discuss experimental design, suggest increasing sample size from 3 mice to 6 in each group(control, treatment), minimizing the system errors
- The experiment will be done in one month
- Will talk with supervisor to discuss the collaboration with our department

18Aug09

William Hardesty, Chemistry; return 1Sep09

- R save() data frame is attached
- Suggest using various of lasso/elastic net with Cox model, forcing age and sex (don't penalize them) into the model
- 62 patients with melanoma
- Comparing protein expression to survival time
- Original approach was to pick a winner by examining separate associations
- Simultaneous modeling with penalization has many advantages; take observed associations with a grain of salt; discounting

- Went through usage of coxpath; best model by AIC or BIC had dozens of features but coefficients may be small
- May be worth looking at coxme function for quadratic penalty

21Jul09

Charles Flynn, Surgery

- Protein and lipidomics pre and post bariatric surgery
    - lean, obese, NASH with 10 patients each
    - ChangYu had estimated power was high; however power estimate was based on the assumption that a single marker to compare (across the 3 groups) was pre-specified
    - Perhaps 3000 candidate lipids to narrow down to one or a few "significant" marks that distinguish the 3 groups
    - May be important to get technical replicates because of variation caused by ions being laid down while the current ions are being processed; but costs $1000/sample
- Lipids extracted by biopsy; rudimentary matching of spectra and proprietary principal components analysis
- Have applied for VICTR funding to supplement a funded grant
- Two DDRC projects funded and pooled
- Aims are quantitative lipidomics and imaging lipidomics; interest in non-alcoholic fatty liver disease
    - a pattern of lipid deposition can promote a specific liver disease

Consultants: LilyWang, NateMercaldo, FrankHarrell

30Jun09

Hernan Correa (Pathology); Maria Piazuelo (Medicine); Pelayo Correa (Medicine, Program Project Grant)

- GI cancer varies greatly from mountain to coastal areas of Columbia
- Results to h. pylori infection modulated by a GI parasitic disease
- Mass spectrometry MALDI-TOF; 90 subjects; low-risk (coast) and high-risk (mountain) groups
- Subjects are males 40-59 years old with symptoms requiring endoscopy
- Surface epithelium is of interest; variation of protein expression over the different histologic types will also be of interest
- Need file with m/z information and a file with clinical information (age, histology,, ...)

- Cancer Center Biostatistics may be able to help if this provides preliminary data for a grant proposal (VICTR can provide support for prelinary results too; not for executing an existing grant)
- This may lead to a line of research in children related to eosinophilic esophagitis (this may qualify for VICTR as a new entity; VICTR application will have to make it extremely clear how this is not primarily related to executing a funded grant)
    - Suggest applying for a $8000 voucher from VICTR and will need to explain how this fits with the already-awarded voucher
- Target is Dec09
- Consultants: Pengcheng Lu, Yu Shyr, Ming Li, William Wu, Steven Chen, Frank Harrell

7Apr09

Ralph Passarella, Undergrad, Molecular and Cellular Biology; working in Radiation Biology

- Submitting a paper on mice tumor response to therapy after a peptide treatment
- Using imaging to get % binding, comparing treated to untreated
- In analyzing response to treatment need to separate binding from vasculature response
- 6 treatment groups; each has n=4 except positive control (n=9) and one technical problem resulting in an n=3. These pertain to plates (cell culture dish). Cell lines were used. Plates represent experimental units.
- Need to display raw data; see DynamitePlots
- Might display as a 3x2 matrix of dot plots or group pairs
- Unified analysis with 6 treatment groups, organized as a two-way ANOVA with a 3x2 setup, is recommended
    - Assumes normality and constant variance across 6 groups; assumes data properly transformed
    - Alternative approach: rank-based analysis; proportional odds ordinal logistic regression can test for interaction and test contrasts
    - Proportional odds model is a generalization of the Wilcoxon/Kruskal-Wallis test and does not require one to properly transform the response variable (% binding); it is also robust to outliers and does not assume normality of the raw data
    - Data layout (e-mail to biostat-clinic@list.vanderbilt.edu): column for celltype, column for treatment, column with response measurement; # rows = total # plates (long and thin spreadsheet)
- Suggested R code

```
library(Design) dd
```

18Dec07

Dan Kaiser - Medical Student 4th Yr

p

13 May 2008

Discussions: Amino acids, pseudocounts and "propensities"

17 Nov 2008

Discussions: M. Chambers-ETD mass spectra charge determination modeling Suggestions: logistic regression (unordered or ordered?); multinomial logistic regression; get a probability for each category? select N most probable charge states instead of just top probability? make sure normalization is not predictive of outcome (charge)

20 Jan 2009

Discussions: Jiajun Shi: Gene Association studies; two-stage confirmation; Winner's Curse: http://www3.interscience.wiley.com/journal/121591995/abstract

"Meta-analysis" of data collected for a different study: can data collected for a control/breast cancer/diabetes study be reused for studying a different phenotype (obesity)? Some paper has already done this, but faculty survey says that's a bad idea because of the variation it would introduce.

24 Feb 2009

Discussion: Britney Grayson's Sample size and power analysis for copy number variation chip. Experiment Design: 10 diseases (Type I diabetes) 10 controls, they are 10 monozygotic twins; >1.8 million spots Suggests: Adapt design by adding samples based on initial results, power calculation for the qPCR outcomes. Come in next week

3 Mar 2009

Cathy Derow: binary outcome install.packages('glmnet') library(glmnet) library(Hmisc) xless(glmnet) y Note: Had to run R as the superuser, otherwise R could not find =glmnet

10 Mar 2009

Issue about limits of knowledge of multiple markers related to a binary endpoint, and futility of finding a cutpoint. library(Design) x1