From Omics to EHR, Following the Research Calls (Needs)

Xiangqin Cui

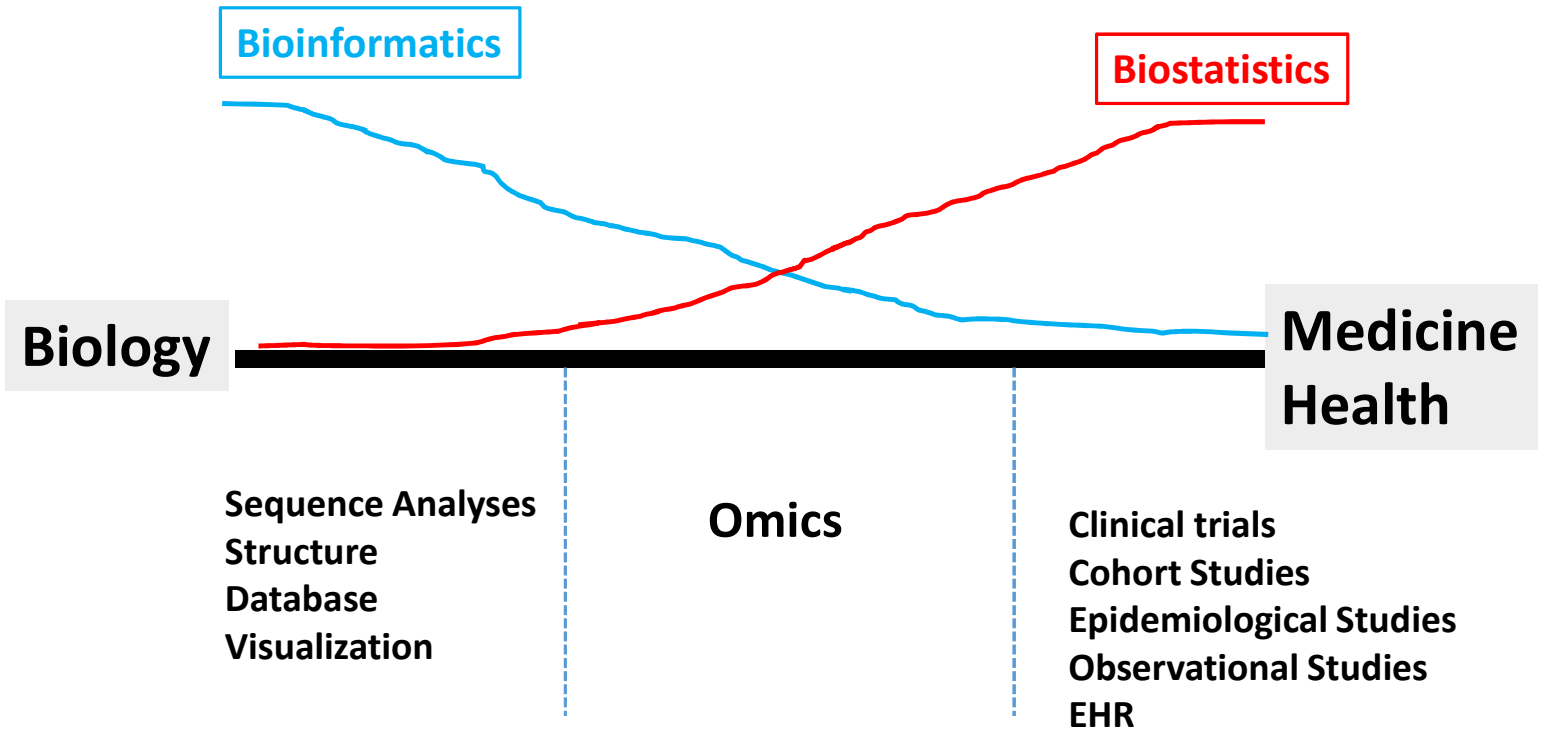
2/28/2020 at SPCC

Outline

- My understanding and practice of Bioinformatics/Biostatistics
- Cancer project on Patient Derived Cancer Model (RNAseq)
- EHR projects
- Thoughts on statistical collaboration

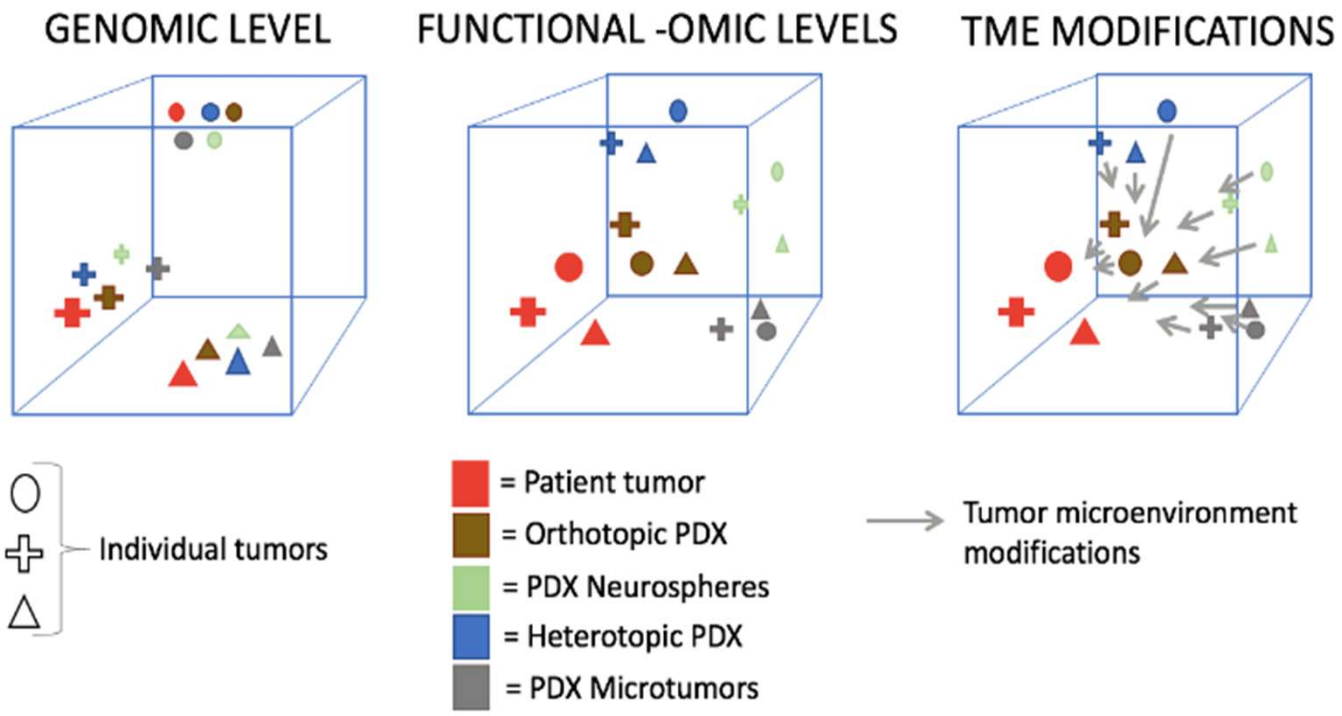
A Little Bit About Me

- 1994 – 2001 PhD in Genetics at Iowa State University
  (Maize Male Sterility Restoration)
- 2001 – 2004 Post Doctoral Training at the Jackson Lab
  (Statistical Genetics/Experimental Design and Data Analysis of Microarray)
- 2004 – 2017 Faculty in Department of Biostatistics at University of Alabama at Birmingham
  (Experimental Design and Data Analysis for microarray, RNAseq, DNA methylation, microbiome, microRNA etc.)
- 2017 – present Faculty in Department of Biostatistics and Bioinformatics at Emory University. Director of Atlanta VA Data Analytics Core.
  (VA EHR Data Extraction, Analysis, intervention trials, etc.)


Bioinformatics
Biostatistics
Biology
Medicine
Health
Sequence Analyses
Structure
Database
Visualization
Omics
Clinical trials
Cohort Studies
Epidemiological Studies
Observational Studies
EHR

7 17 3

U01 – Biological Comparisons in Patient-Derived Models of Cancer (GBM)

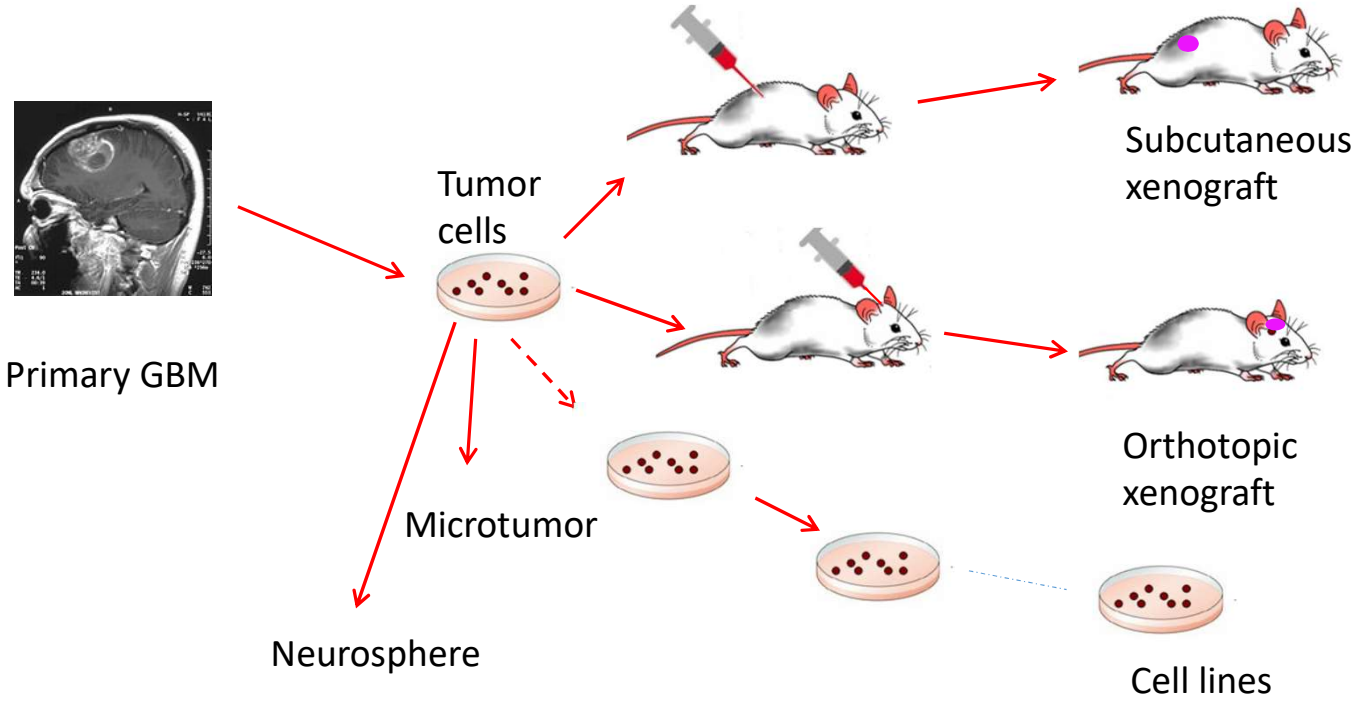


My Job in the U01

- To identify a good similarity metric for clustering and comparing samples from patients and derived models.
- Equivalence/similarity analysis at gene level and kinase level.

GBM Models

- GBM Model building for drug development and mechanism studies



Fidelity of Changes in Different GBM models

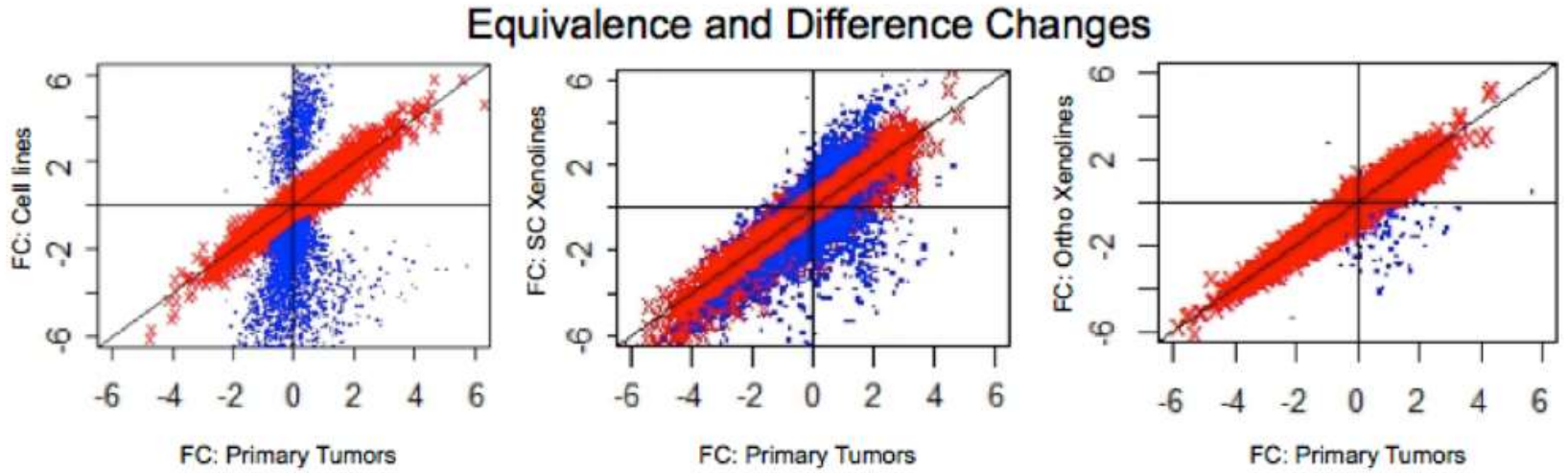
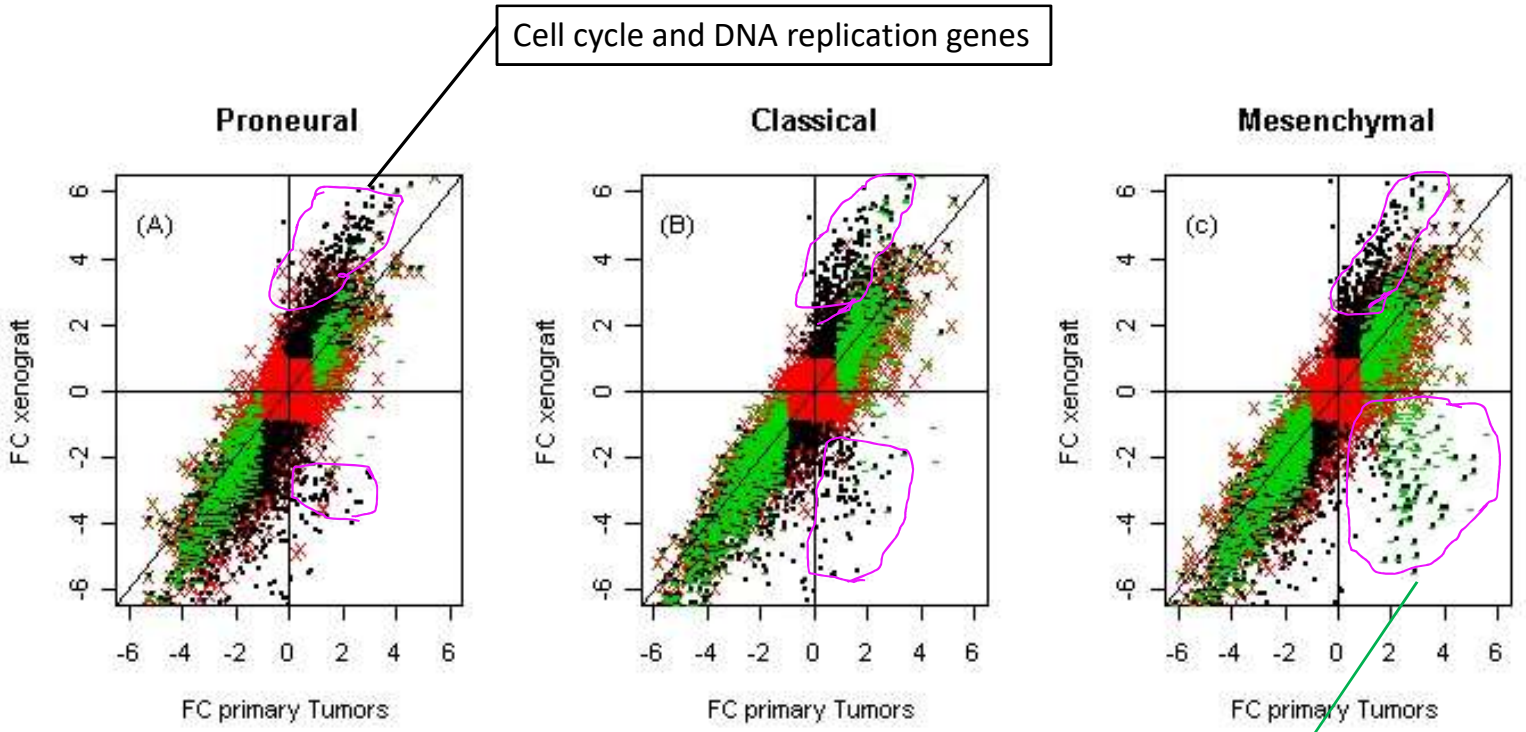


Figure. 1. Gene Expression comparison of Glioma Cell Lines (Left), Subcutaneous xenolines (middle) and Orthotopic Xenolines (right) to the GBM primary tumor (abscissa). Probe sets that were deemed to be significantly different at FDR 0.05 and fold change greater than 2 (blue points) are shown relative to the equivalent (at FDR 0.05) probesets (red points).

Fidelity in Different GBM Subtypes



Red: equivalent changes
Green: Significant changes in primary GBM
Black: Significant changes in Xenografts

Similarity Metrics

- 46 traditional distances in R package “philentropy”
- traditional distances:
  - Euclidean
  - Pearson correlation
  - Neyman
- Novel distances
  - ISC (Sohangir & Wang, 2017),
  - Ahmad(Hassanat & Hassanat, 2014),
  - EucPear (Yona, Dirks, Rahman, & Lin, 2006),
  - MD (Yona et al., 2006) (eliminated later)

Data and Genes for Evaluating similarity metrics

We generated 2 datasets: 7-patient set and 9-patient set.

- 7-patient set (25 samples)
  - 3 normal patients (IDs: 1034, 2004, 2060)
  - 4 patient tumors and their Xenograft samples (IDs, 1011, 1016, 1046, 1060).
- 9-patient set (35 samples)
  - 3 normal patients (1034, 2004, 2060)
  - 6 patient tumors and their Xenograft samples

❑Genes

- ❑ microarray with 22011 genes
- ❑763 signature genes for sub-classifying tumors

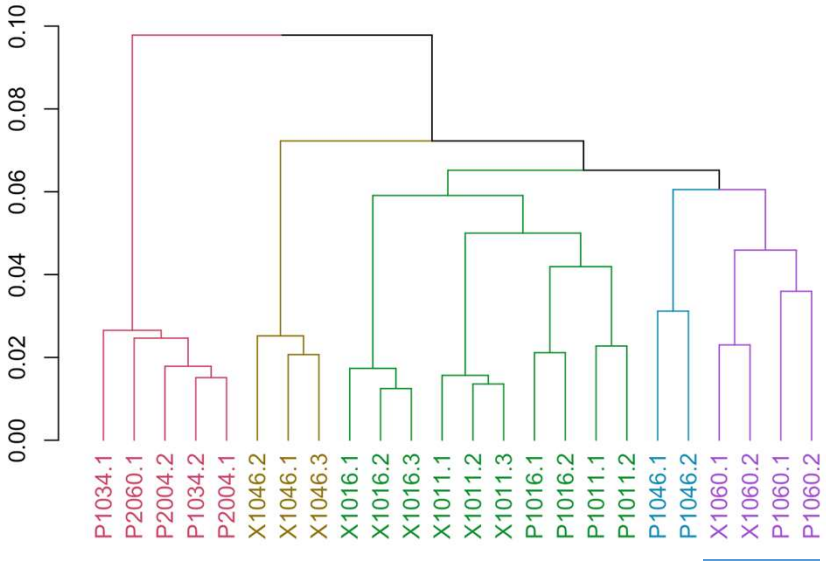
Data preprocessing

- Deconvolution to remove mouse contamination in the xenograft samples
- Normalization: Quantile Normalization
- Filtering low expressers <7 on log2 scale

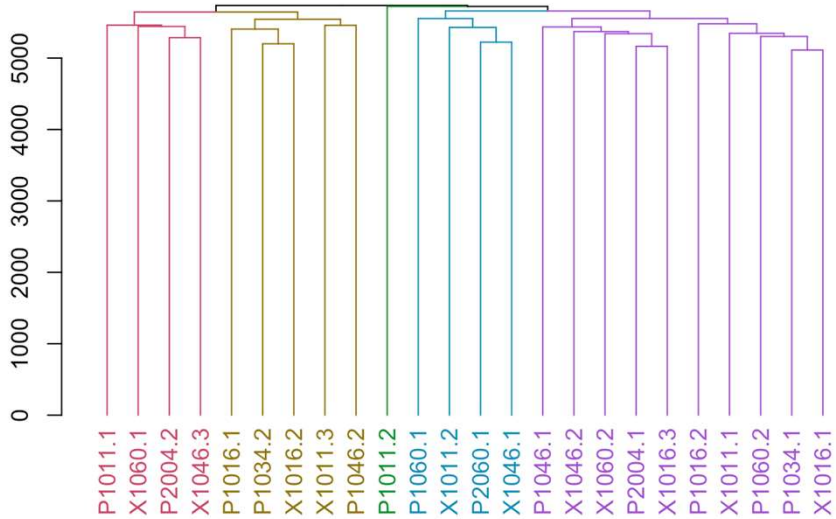
- hierarchical clustering under several situations:
  - with/without deconvolution,
  - with/without normalization,
  - using all genes/only using signature genes.
- We found that the some distances give similar results including manhattan, neyman, Sorensen, etc. However, they mostly cluster patient samples together and xenoline samples together and occasionally could correctly cluster paired patient sample and xenoline samples together.

Example of some clustering results

Manhattan distance



intersection distance



In combination with clustering methods

	Hierarchical		k-medoids		kmeans	
	gene > 7	signature genes	gene > 7	signature genes	gene > 7	signature genes
euclidean	0.881	0.886	0.879	0.884	0.854	0.815
EucPear	0.881	0.886	0.879	0.884	0.854	0.834
ISC	0.882	0.886	0.879	0.866	0.854	0.827
neyman	0.882	0.877	0.884	0.872	0.842	0.835
Pearson	0.881	0.886	0.884	0.884	0.854	0.84

Evaluated with Rand Index (RI)

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Think about the process/analysis of omics data.

Example: RNA-seq analysis

Experimental design

- Statistical experimental design principles
  - Randomization
  - Replication (small sample size)
- Sequencing depth
- Biases of NGS

Fang and Cui (2011) Briefings in Bioinformatics, 12:280-7

Preprocessing of RNAseq sequences

- Illumina software generates **fastq** files (sequence + quality score)
- Quality control of sequences (e.g. FastQC)
- Alignment (e.g. STAR)
- Get a table of read counts (one count per gene per sample) (e.g. HTseq)

NGS biases

- Transcript length and sequence depth
- GC content
- Other sequence compositions

Methods for Adjusting Length and Total Reads

- Upper-quartile (Bullard et al., 2010): using not the total read but the upper-quartile
- TMM (Robinson and Oshlack, 2010): Trimmed Mean of M values
- Quantile (Irizarry et al 2003)
- FPKM (Trapnell et al., 2010): Cufflinks software Fragments Per Kilobase of exon per Million Fragments mapped.
- FVKM (fragments per virtual kb per million) (Lee et al 2011) based on common region or isoform unique region.

modified from Somia Tarazona presentation at http://www.slideshare.net/cursoNGS/sonia-tarazona-differential-expression

Generalized Additive Model Based Algorithm to Jointly Correct for GC Content Bias and Sequence Bias

For each gene:

$$Y = \alpha + X_{length} + X_{AA} + X_{AT} + X_{AG} + X_{AC} + ... + X_{CC} + \varepsilon$$

reads

Expression

Length bias

Dinucleotide bias

(obtain PCs that explain 95% variation)
Then replace the dinucleotides.

generalized additive model (GAM):

$$log(Y_i) = \alpha_i + X_{i,\,1} + s(P_{i1}) + s(P_{i2}) + ... + s(P_{iK}) + \varepsilon_i.$$

Zheng et al (2011) BMC Bioinformatics. 12:290

Test for Differential Expression (with biological replicates)

- Due to the over dispersion of variance, Negative Binomial is used for modeling the gene level reads.

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)$$

Gene i
Sample j

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}$$

Estimate as a smooth function of expr at each gene

$$v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)})$$

DEseq Package

Anders and Huber *Genome Biology* 2010, **11**:R106

Bayesian Hierarchical Models

- At noise level
    - $K \sim \text{poisson}(\lambda)$ (shot noise)
    - $\lambda \sim \text{Gamma}(r, p/(1-p))$ (biological variability)

Clustering and Classification

- Feature selection
- Unsupervised clustering (hierarchical, k mean)
- Superverse classification (regressions, machine learning)

Omics vs HER (Data characteristics)

	Omics	EHR
Study	Designed	Salvage from existing records
Cost	High	Low
Data quality	High	Low
Sample size	Small (except GWAS)	Huge
Missing	More systematic	Every where, irregular, not random
Messiness	Neat	Messy

Omics vs HER (Data Access and Analysis)

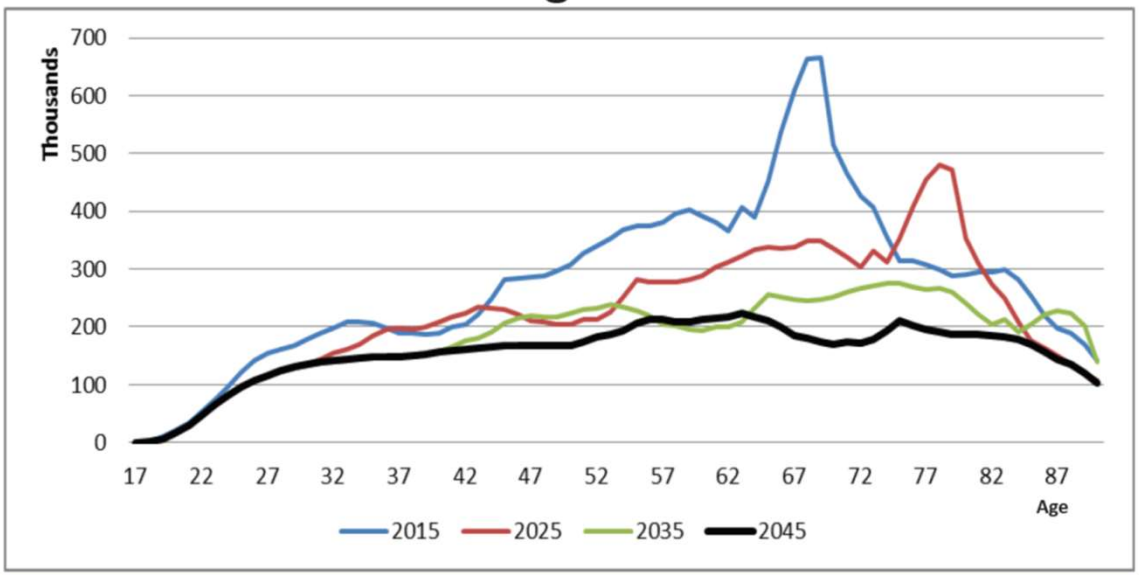
	Omics	HER
Data Access	Easy	Much more difficult
Data Extraction	streamlined	Complex process for every project (ICD codes, NLP)
Data Cleaning	Diagnosis (keep, remove)	One variable at a time (Multiple rounds)
Analysis	Packages, Methods	Classic methods, machine learning
Missing Data	Imputation	Simple methods, e.g 2-year average
Major Concerns	Multiple testing, batch effect	Super small p vales very small effect size, bias.

Electronic Health Records (EHR) Data

VA EHR – CDW (Corporate Data Warehouse)

- **More than 9 Million current patients, 172 medical centers in 21 VISNs.**
- **It has 20 years and 23.5 million patients in production database (since 1999).**
- **Contents: Appointment, Lab, Allergy, Consult&procedure, Orders, Health Factors, Immunization, Inpatient, Mental Health, Out patient, Primary care Management, Pharmacy, Purchased Care, Surgery.**
- **VA Informatics & Computing Infrastructure (VINCI) converts into searchable data in relational tables.**
- **Updates daily.**

Figure 1.
Veteran Age Trends

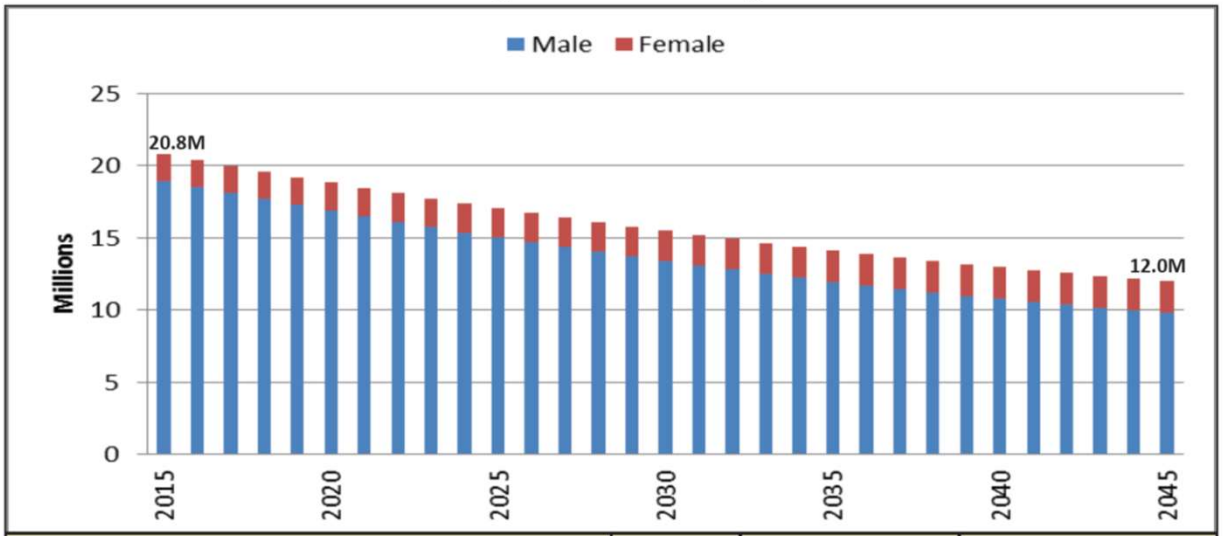


Over the next 30 years Veteran age becomes more evenly distributed

Source: Veteran Population Projection Model 2016: Congressional Briefing, June 2017, revised. Prepared by the National Center for Veterans Analysis and Statistics as of May 14, 2018.

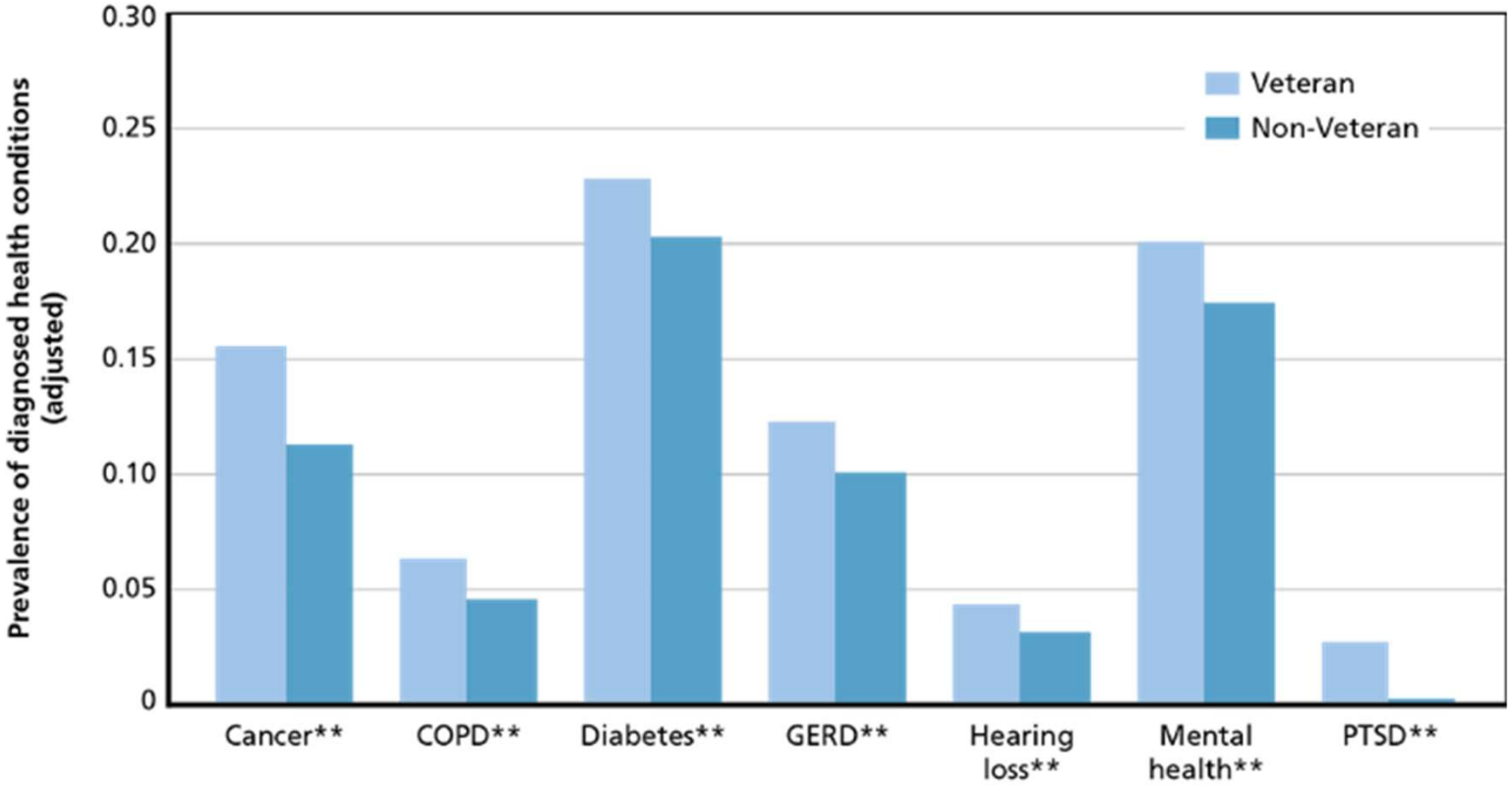
https://www.va.gov/vetdata/docs/Demographics/VetPop_Infographic_2019.pdf

Veteran Projections by Gender



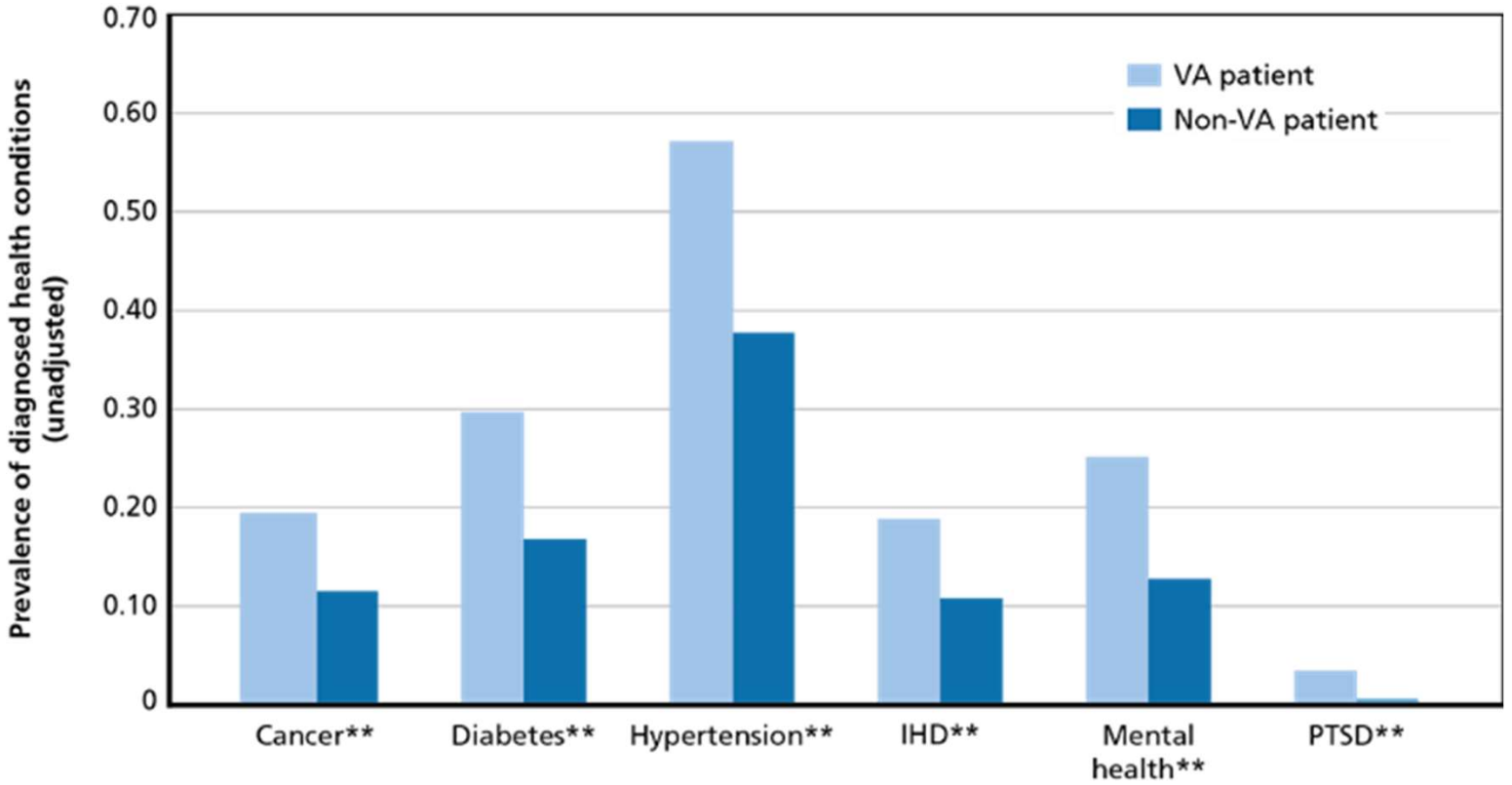
Annual % Change of Veteran Population Over 30 Years	Total	Male	Female
VetPop2016	-1.8%	-2.2% (Blue)	+0.6% (Red)

Source: Veteran Population Projection Model 2016: Congressional Briefing, June 2017, revised. Prepared by the National Center for Veterans Analysis and Statistics as of May 14, 2018.



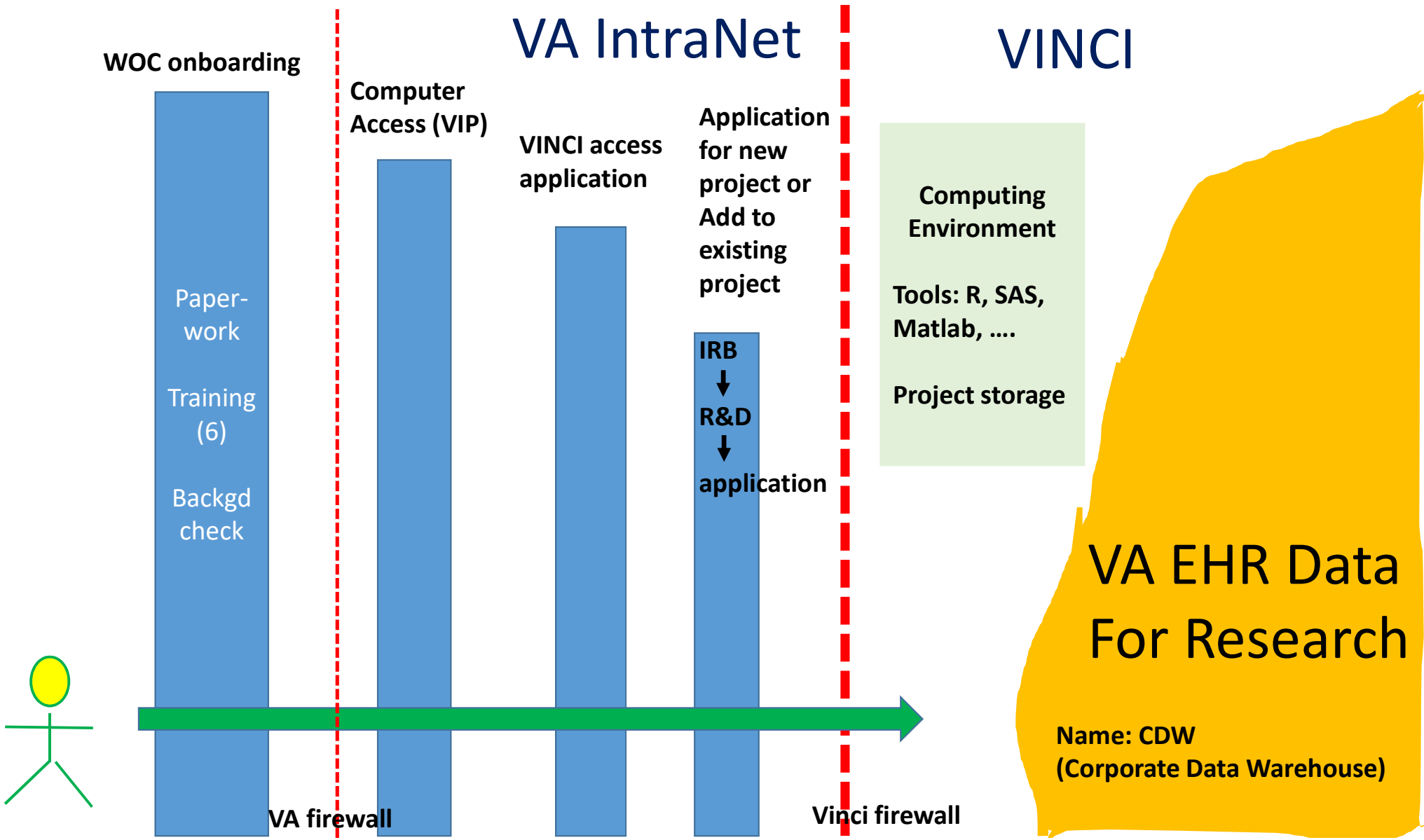
SOURCE: RAND analysis of MEPS, 2006–2012.
NOTES: ** indicates a statistically significant difference between Veterans and non-Veterans at p-value < 0.05. Sample size, non-Veterans = 150,225, and sample size, Veterans = 12,313. Sample sizes may be smaller for some conditions due to missing values. The prevalence rate of each health condition is the predicted prevalence in 2014 for the populations of Veterans and non-Veterans, both with age, sex, race/ethnicity, region, and urbanicity, adjusted to match the demographic composition of Veterans in 2012. Cancer includes any malignancy, and Mental Health includes any mental health condition.

https://www.rand.org/pubs/periodicals/health-quarterly/issues/v5/n4/13.html



SOURCE: RAND analysis of MEPS, 2006–2012.
NOTES: ** indicates a statistically significant difference between VA patients and Veterans who are not VA patients at p-value < 0.05. Sample size, VA patients = 4,871, and sample size, non-VA patients = 7,442. Sample sizes may be smaller for some conditions due to missing values. Cancer includes any malignancy, and Mental Health includes any mental health condition.

https://www.rand.org/pubs/periodicals/health-quarterly/issues/v5/n4/13.html


VA IntraNet
VINCI
WOC onboarding
Paper-work
Training (6)
Backgd check
Computer Access (VIP)
VINCI access application
Application for new project or Add to existing project
IRB
R&D
application
Computing Environment
Tools: R, SAS, Matlab, ….
Project storage
VA EHR Data For Research
Name: CDW (Corporate Data Warehouse)
VA firewall
Vinci firewall

Omics Data Access (in contrast)

- **Download from facility server**
- **Public data freely available such as GEO**
- **dbGaP (genotype and phenotype)**
  Submit an application

Types of projects our Data Analytics Core works on

EHR projects:

- **Knee replacement: N = 12,600 for evaluating center efficiently, pain management, drug usage, complications.**
- **VA-ADPKD cohort building (~ 6000 patients)**

Existing Quality Improvement Programs

- **Empowering Veteran Program (EVP): Behavioral intervention program – 10 wk intervention and follow up. N = 800 on going. Use EHR to find matched controls and similar programs for comparison.**

Intervention Trials: (100-200 subjects) in Rehab and Tele-medicine

Total Knee Replacement

(Drs, Allison Arensman and Blake Anderson)

Objectives:

- Examine the variation in procedure time across the VA healthcare system.
- Identify potential factors influencing procedure time.
- Examine health outcomes influenced by variation in procedure time.

EHR Data Time Span: 2010-2014

Number of patients extracted from CDW : 12,600

Variables Extracted

CPT Code indicating Total Knee arthroplasty; Other CPT codes (if any to determine return to OR in 365 days following index procedure)

- Demographics: age, sex, race/ethnicity, zipcode; Patient BMI; Charlson comorbidity index
- VA Facility
- Surgery variables: Surgical Date; Surgeon; Length of surgery; Involvement of a resident; Presence of tourniquet; Length of tourniquet time if applicable; Type of anesthesia; ASA score; Whether or not an intraoperative xray was taken
- BEER Drugs (categorized drugs that are routinely used such as aspirin.

Outcomes

- Length of Stay (LOS) in days
- Major Complications

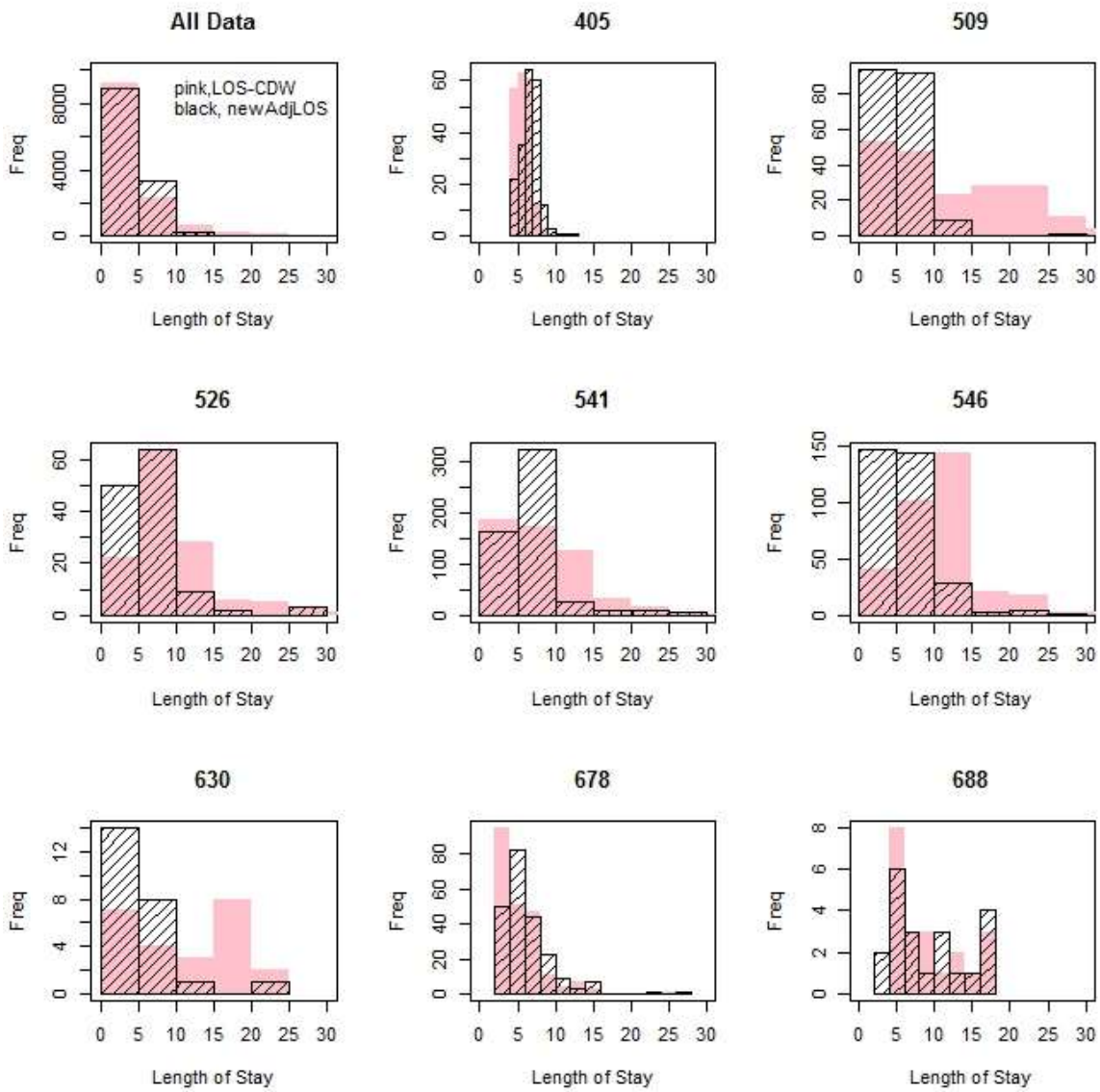
  readmission; reoperation, mortality, ER visits
- Other complication

  Inpatwound; outpatwound; inpatstroke; outpatstroke; inpatpe;outpatpe; inpatpna; outpatpna; inpatdvt; outpatdvt; inpatacutemi; outpatacutemi; inpatsepsis; outpatsepsis; inpataki; outpataki; inpatrespfailure; outpatrespfailure

- Simple analysis methods: linear regression, logistic regression
- Complicated data cleaning

Data Issue Example

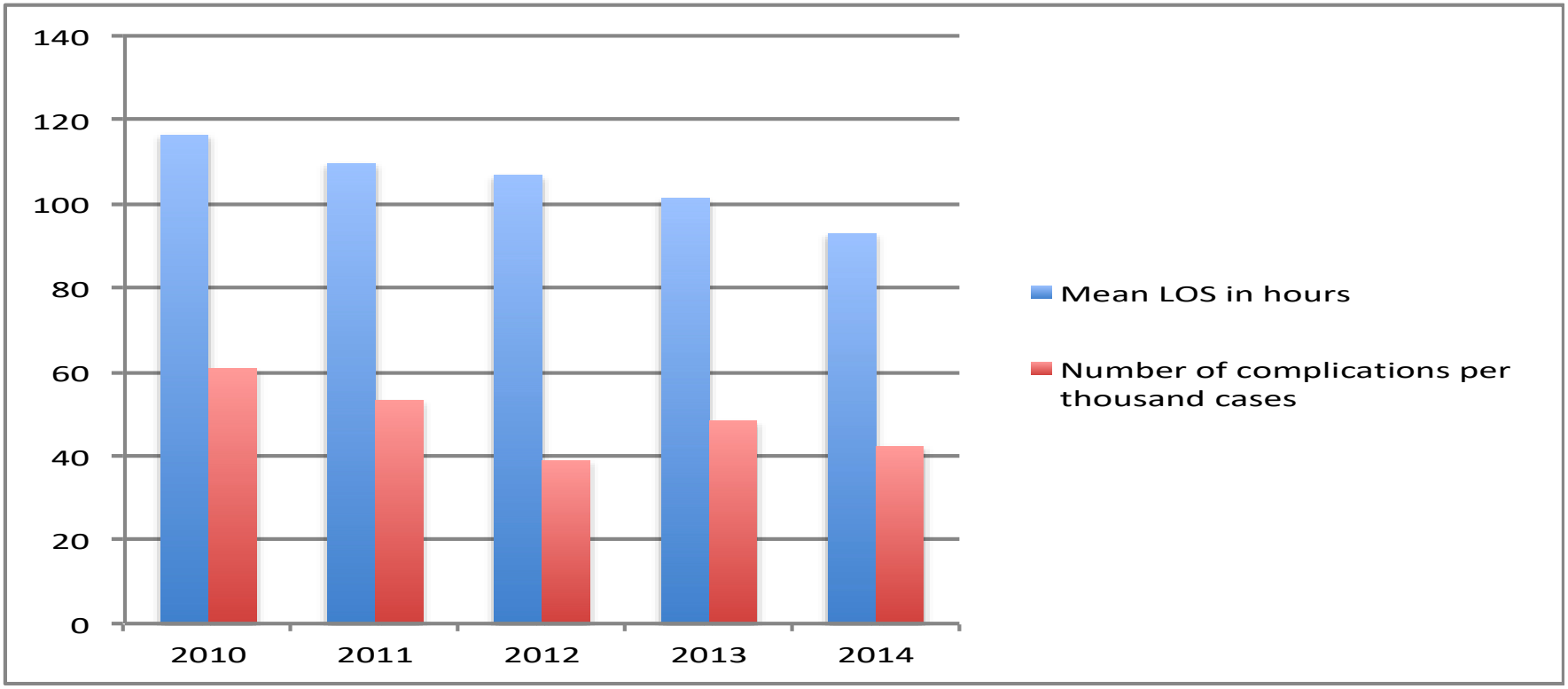
Length of Stay (LOS)
for knee replacement



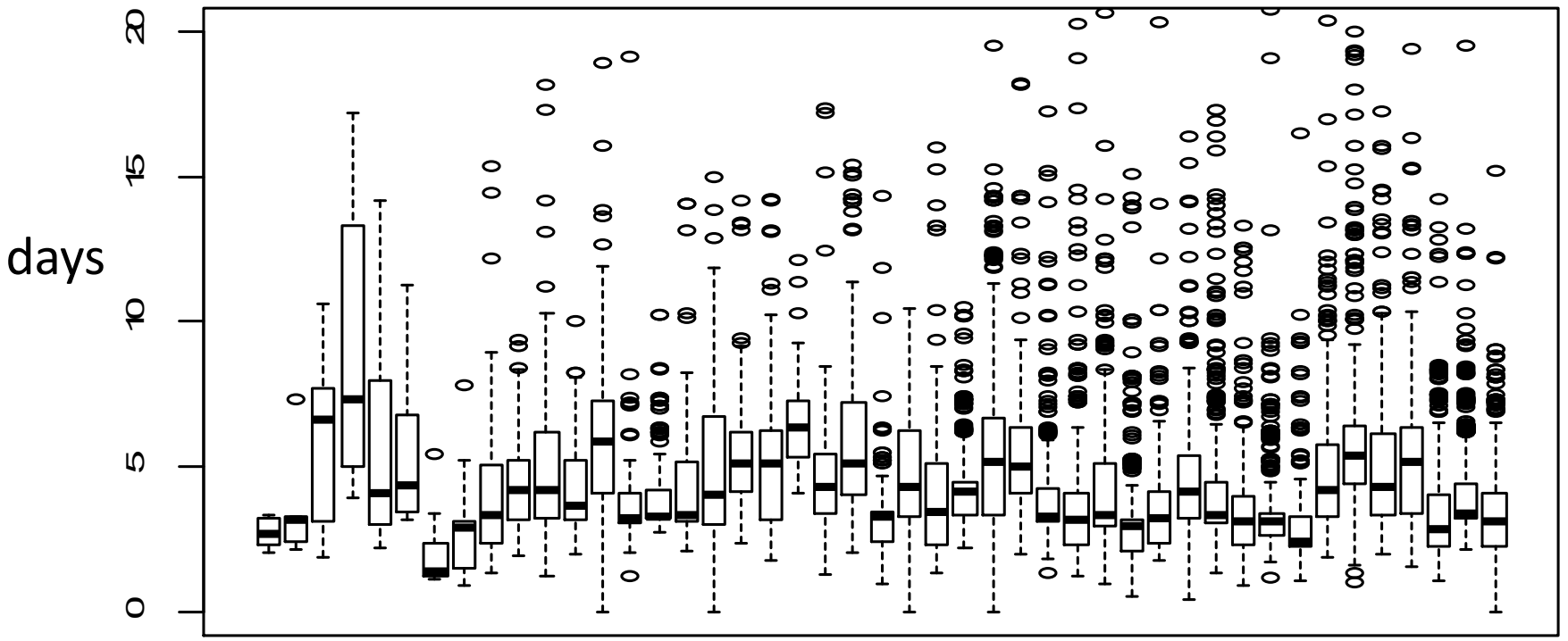
Other data problems encountered

- Length of Stay ~ wrong time stamp or no time stamp
- Race ~ different entries at different time
- Operation time ~ supper short or supper long
- Outliers for every variable
- Data cleaning (5-10 rounds)

Overall Length of Stay over time



VA Length of Stay for Total Knee Replacement

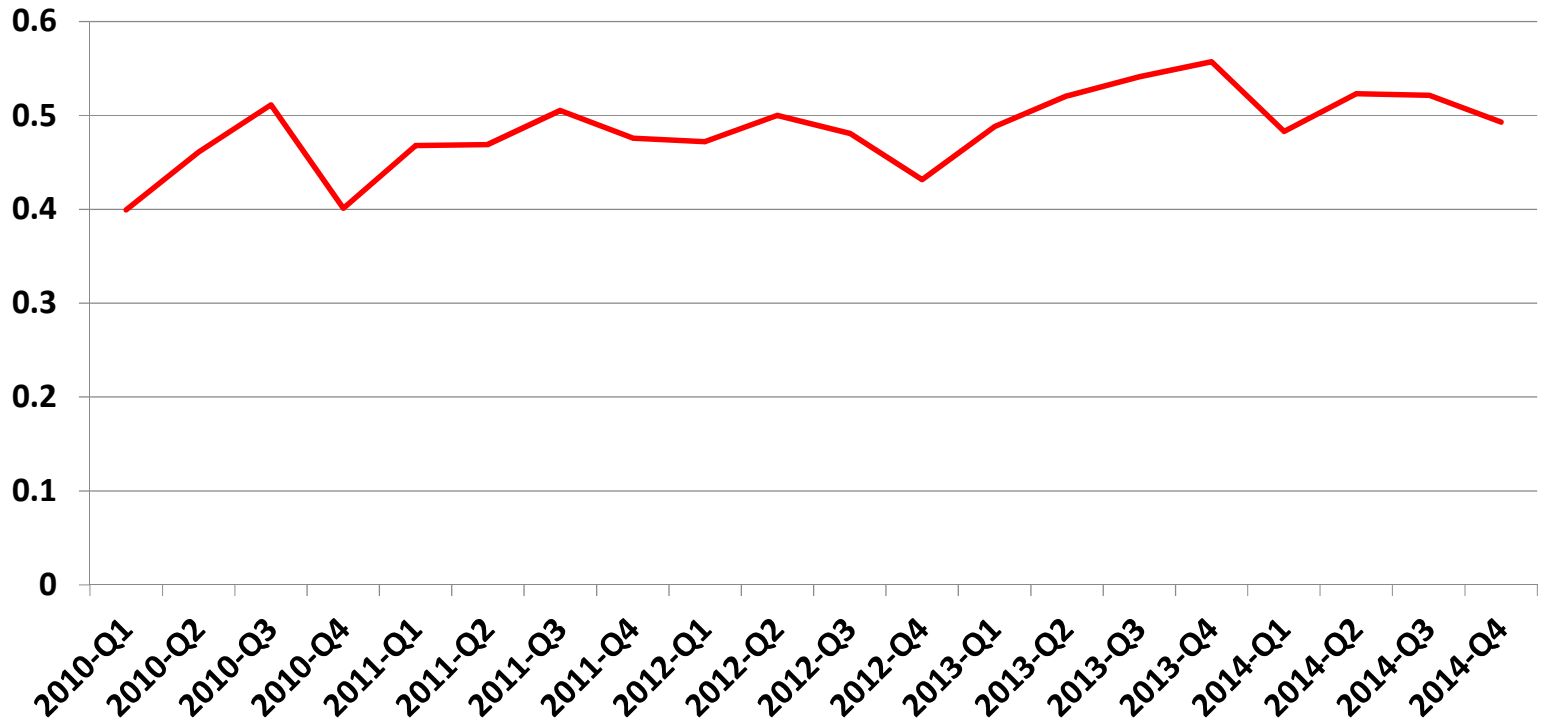


Centers ordered from low to high volume

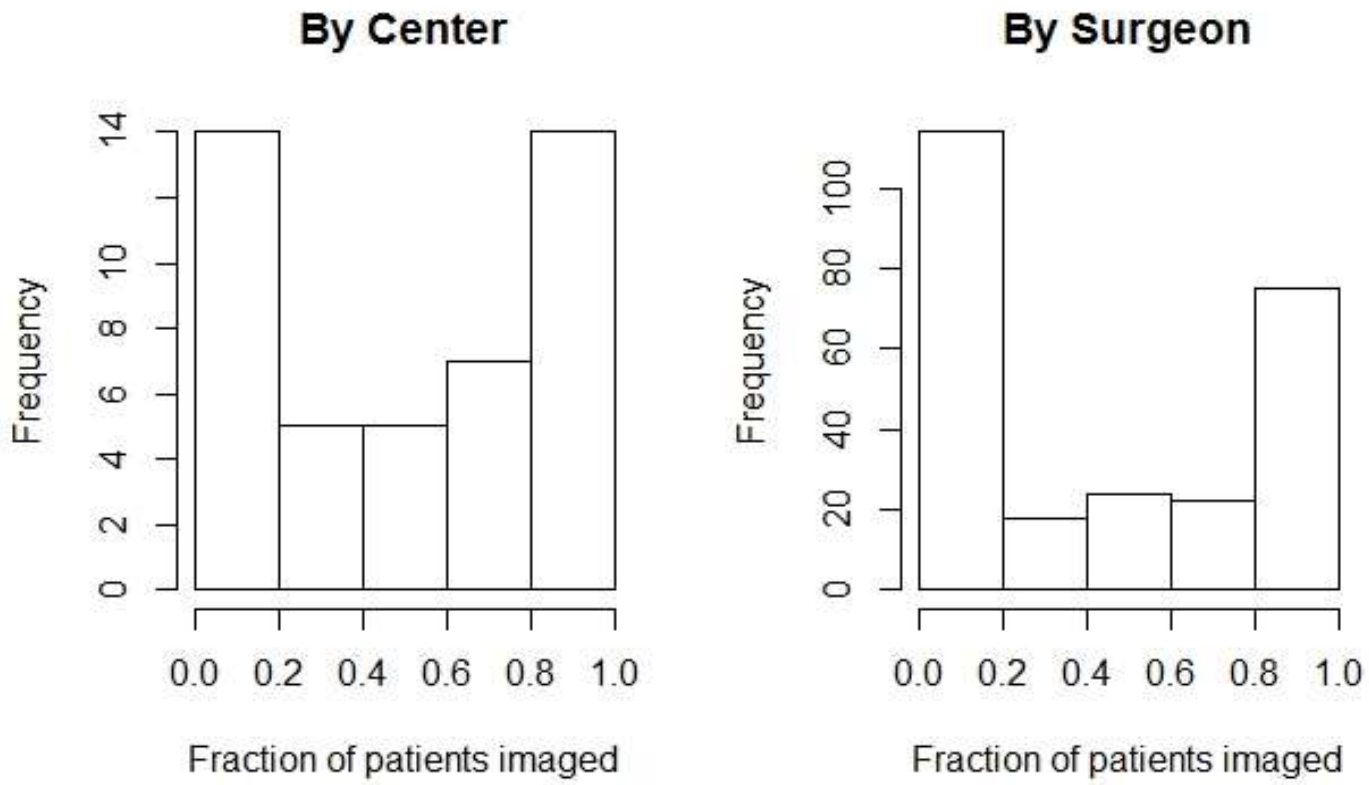
Factors Associated with Increased Length of Stay

Group		Coefficient	Standard Error	p value	R2 (%)
Facility		----	----	<0.0001	16.7
Gender				<0.0001	
	Female	0.11	0.018	<0.0001	0.23
Age		0.01	0.0006	<0.0001	2.1
BMI		0.0028	0.0009	0.0014	0.015
ASA Score		0.063	0.001	<0.0001	0.25
Requested Anesthesia Type				0.021	0.66
	Block	0.037	0.038	0.33	
	Central	-0.034	0.013	0.011	
	Spinal	-0.036	0.016	0.025	
	MAC	-0.14	0.1	0.17	
Race				<0.0001	0.43
	Black	0.1	0.013	<0.0001	
	Other	-0.014	0.016	0.39	
Charlson Score		0.026	0.0029	<0.0001	0.5

Fraction of patients imaged during TKA operation



Frequency of Imaging by Center and by Surgeon



Factors Associated with 30-day Major Complications			
	Odds Ratio	Confidence Interval	P-Value
Gender (Female)	0.9114	(0.64-1.30)	0.6098
Age	1.0216	(1.01-1.03)	<0.0001
Race (Black)	1.5009	(1.20-1.88)	0.0004
Race (Other)	0.7348	(0.52-1.04)	0.0862
BMI	1.0095	(0.99-1.03)	0.2609
ASA Score	1.0768	(0.88-1.31)	0.4594
Charlson Score	1.1001	(1.05-1.15)	<0.0001
Intraoperative X-Ray	1.1394	(0.89-1.45)	0.2937

Omics vs HER (Data characteristics)

	Omics	EHR
Study	Designed	Salvage from existing records
Cost	High	Low
Data quality	High	Low
Sample size	Small (except GWAS)	Huge
Missing	More systematic	Every where, irregular, not random
Messiness	Neat	Messy

Omics vs HER (Data Access and Analysis)

	Omics	EHR
Data Access	Easy	Much more difficult
Data Extraction	streamlined	Complex process for every project (ICD codes, NLP)
Data Cleaning	Diagnosis (keep, remove)	One variable at a time (Multiple rounds)
Analysis	Packages, Methods	Classic methods, machine learning
Missing Data	Imputation	Simple methods, e.g 2-year average
Major Concerns	Multiple testing	Super small p vales very small effect size, bias.

Challenges of Statistical Collaboration

- Knowledge gap between statistician and investigator (domain expert).
- Who fills the gap or builds the bridge?
- How to prioritize which research area to build the bridge for?
- What about all the other requests for services?

VA-PKD Cohort (Progression Prediction)

- **Estimated ~6000 PKD patients in the VA CDW.**
- **We plan to use them for predicting disease progression.**
- **Data issues:**
  - **eGFR (outcome) has two values for each person based on race in some centers and one value reported from others.**
  - **“>60” instead of value**
  - **Medication is massive**
  - **Irregular visits**
  - **Inconsistent reports depending on phisicians**

Acknowledgements

GBM PDX project

- Emory
  - Tianyu Zhang
  - Roshan Darji
- UAB
  - Christopher Willey, MD, PhD
  - Yancey Gillespie, PhD
  - Anita Hijelmeland, PhD
  - Jake Chen, PhD

VA Knee Replacement Project

- Emory
  - Mofei Liu
  - Shiyu Chen
- Atlanta VA
  - Allison Arensman, MD
  - Blake Anderson, MD