A Method for Simultaneous Batch Effect Correction and Analysis of Metabolomics Data

In the Absence of Internal Standards

Stephen Salerno

Department of Biostatistics · University of Michigan

March 14, 2018

Disease Burden of Tuberculosis

Tuberculosis (TB) is the leading cause of death due to an infectious agent:

- Airborne transmission of TB from host to host
- Increased spread of multi-drug resistant Mycobacterium TB
- Early diagnosis is essential for control and treatment of TB

New Cases	Missed Cases	Deaths
9.6 Million / Year	3.5 Million / Year	1.5 Million / Year
26,000 / Day	9,000 / Year	4,100 / Day

Issues in Diagnostic/Prognostic Testing

A limitation of current diagnostics is the need for sputum-based assays:

- Sputum microscopy and microbiologic culture are the gold standard:
  - Varying sensitivity according to collection and processing techniques
  - Cultures require 3-6 weeks of incubation
- Nucleic acid amplification-based methods (GeneXpert):
  - Expensive in resource-limited areas where disease burden is highest
  - Cannot detect extra-pulmonary disease with fidelity

Looking to Metabolism for Answers

Metabolomics involves the simultaneous analysis of hundreds of small molecule compounds, or metabolites, in biological systems:

- Can provide direct biochemical readouts of cellular and organismal behavior and lead to biological insights
- Quantitation of cellular metabolites can be measured using high-throughput techniques including Mass Spectrometry (MS)
- Applications of metabolomics is a growing area of research from basic biochemistry to human health and disease

Metabolomics and TB

Metabolite signatures present in urine may be associated with TB infection or response to TB treatment:

- TB might be associated with urinary metabolic biomarkers for diagnostic testing and prognostication
- Weill Cornell study reported/validated metabolite bio-signatures unique to patients with active pulmonary TB
- Certain statistical considerations make analyzing metabolomics an interesting data problem

Batch Effects in LC-MS Data

Interpretation of metabolomics data is limited by appropriate mathematical tools for normalization and downstream data processing:

- Cross-study comparisons and meta-analyses are currently impractical due to the unknown experimental, technical, and biological variability
- Batch Effects: All undesirable variation in data collected by different operators in different facilities and at different time points
- Sources of batch effects include:
  - Differences in instrument performance / the state of the LC column
  - Differences in preparation of batches sample handling
  - Many other unmeasurable environmental and technical factors

Current Statistical Methodology

Several methodologies are currently available for analyzing metabolite profiles, both with and without controlling for potential batch effects:

- Many methods use heavy isotope spike-in quality controls (QCs) or direct measurement of batch factors and/or injection times
- Normalization methods work by performing variants of Principal Component Analysis (PCA) on the QC data
- Others have proposed clustering-based signal drift algorithms using reference samples to correct for batch effects

Current Statistical Methodology

The main drawback to the use of such techniques remains the necessity of prior information which may, or may not, be available to researchers

- Internal controls and reference samples have practical limitations
- Normalization often does not factor in differences in signal intensity distributions or feature drift patterns, which are experiment-specific

There is a need for mathematical methods that do not rely on such internal controls to analyze metabolomics data

RRmix Model

We developed the RRmix model with the goal of capturing unobserved variation through the inclusion of latent factors. It is defined as follows:

$$y_g|\beta_g, F_g, \sigma_g^2 = \mu + X\beta_g + X_c\gamma_g + \Lambda F_g + W_g$$

with $\sigma_g^2 \sim IG(A, B)$; $W_g|\sigma_g^2 \sim N_n(0, \sigma_g^2 I_n)$; $F_g \sim N_q(0, I_q)$; $b_g \sim Bern(p)$;

$$\beta_g|b_g \sim N_2\left(\begin{bmatrix} 0 \\ b_g\psi \end{bmatrix}, (1-b_g)\begin{bmatrix} \sigma_0^2 & 0 \\ 0 & 0 \end{bmatrix} + b_g\begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}\right)$$

The posterior mean of b_g for each compound g is the posterior probability that g is differentially abundant between groups.

Study Objectives

In order to best analyze the metabolomic TB data, we developed this novel algorithm and compared its performance to other known methods to:

- Show that linear mixed effects models produce satisfactory results in the presence of batch effects
- Introduce an algorithm - RRmix - for differential abundance analysis using metabolomics data
- Demonstrate the feasibility of systematically standardizing data without internal controls or prior knowledge of technical variation

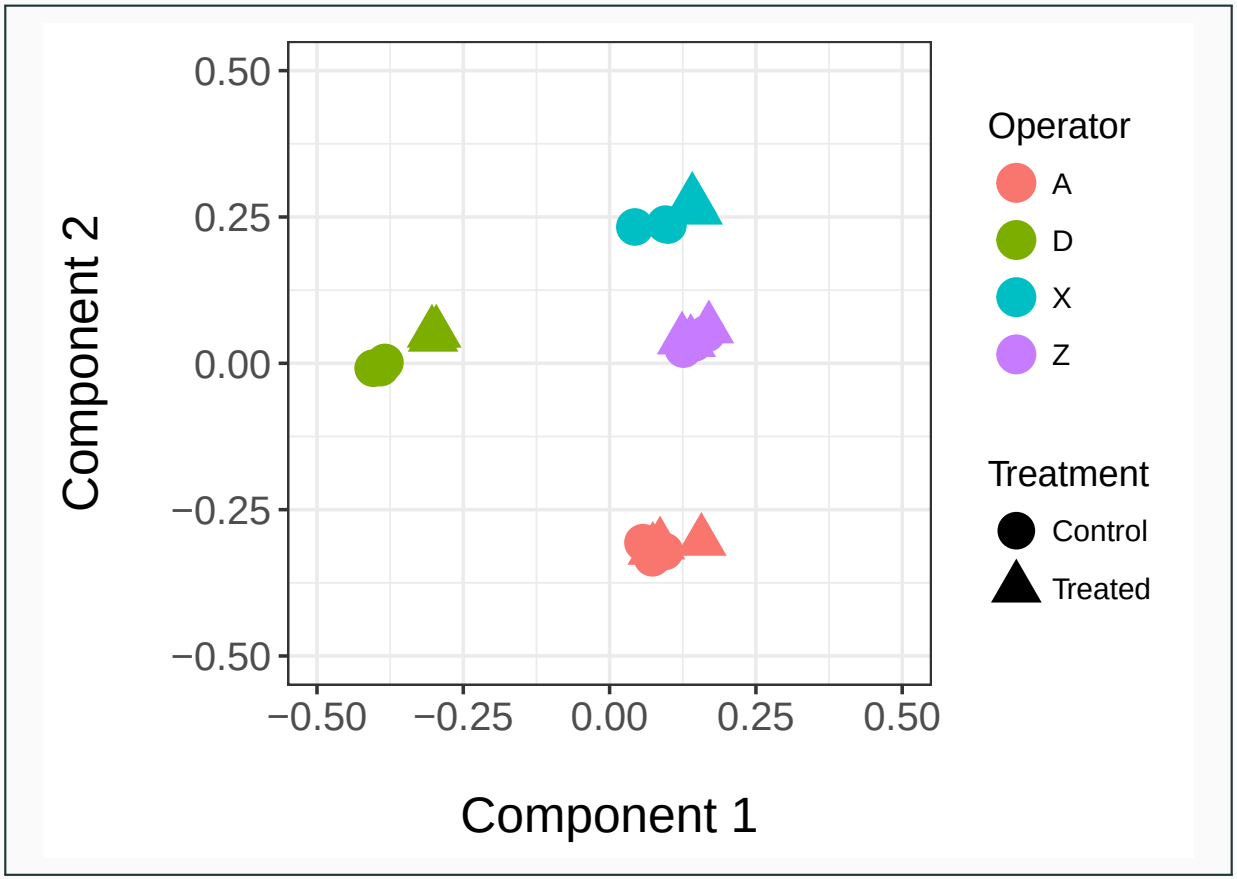
Cell Culture and Drug Treatment

Metabolite samples used for metabolomics were derived from HCT116 colorectal cancer cell lines. Cells were:

- Grown in RPMI 1640, 10% fetal bovine serum, 100 U/mL penicillin, and 100 mg/mL streptomycin
- Cultured in a 37°C, 5% CO2 atmosphere
- Seeded at a density of 3x105 in a 6-well plate and allowed to grow to 80% confluence for each extraction experiment
- Washed with phosphate buffered saline (PBS) and treated with either 5mM of 2-deoxy-D-glucose (2DG) (Sigma) or 0.01% DMSO (cellgro) for 6 hours

Controlled Experiment

Four operators performed the LC-MS experiment in triplicates (3 control samples and 3 treated samples, $n = 24$ total):

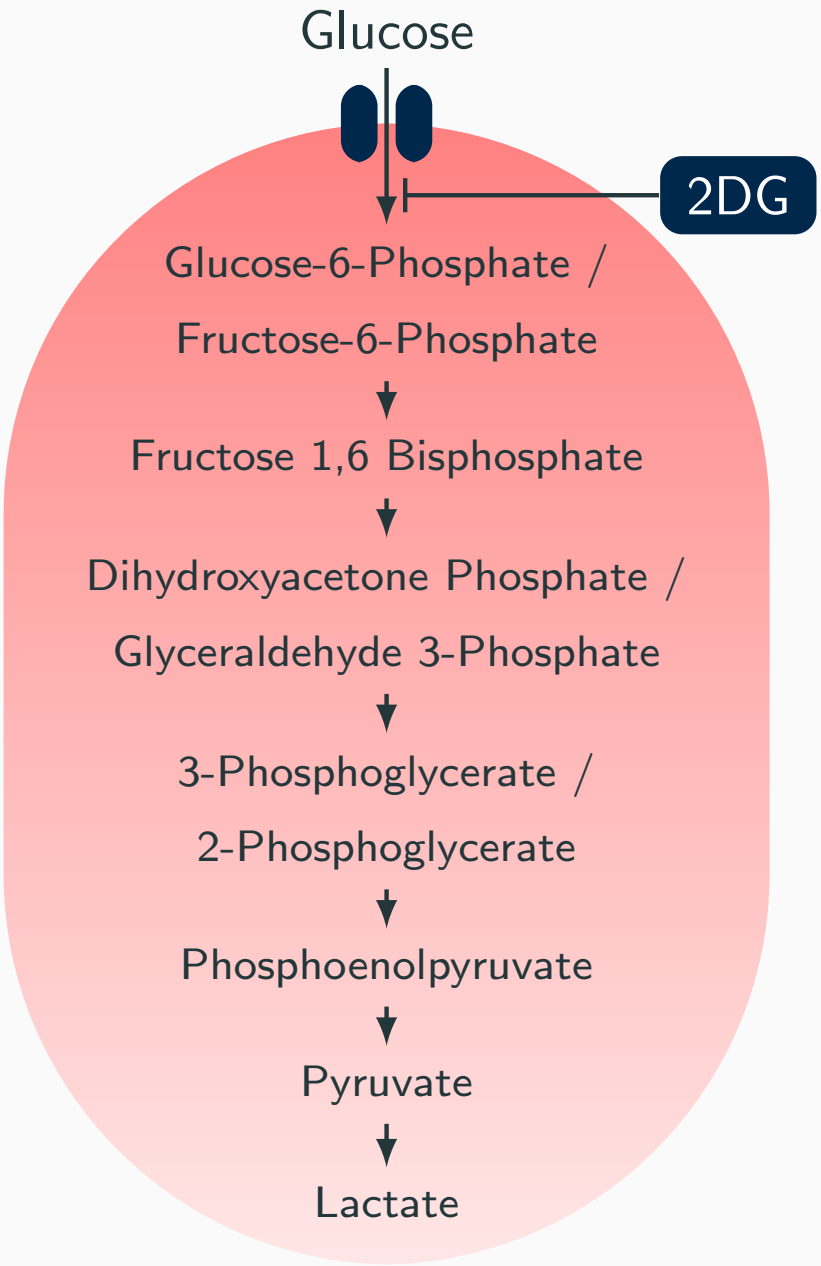


Principal Component Analysis of LC-MS Samples

Positive Controls Along the Glycolysis Pathway

Schematic depicts enzymatic steps in glycolysis pathway and step where 2DG inhibits upon treatment

The metabolites shown in the diagram were analyzed by LC-MS as positive control compounds



Analysis of Controlled Experiment

We used four methods for the analysis of these data (265 metabolites), in presence of an operator-specific batch effect (12 samples):

- Individual t-Tests
- Linear Models for MicroArray Data (LIMMA)
- Factor Analysis model for Multiple Testing (FAMT)
- Random main effect and Random compound-specific error variance model with a mixture structure (RRmix)

Linear Models for MicroArray Data (LIMMA)

LIMMA does not account for latent variation, but it does have several key properties for the analysis of high-dimensional biological data:

- Calculates a moderated t-statistic, $\tilde{t}$, which uses a shrinkage estimate of the standard error in the denominator of the t-statistic
- $\tilde{t}$ is more robust to small metabolite-specific sample variance estimation than t.
- Involves closed-form estimates of the hierarchical model parameters
- Implementation of this modeling strategy is well-documented in a Bioconductor package for `R`

PCA-LIMMA and SVA-LIMMA

Packages such as LIMMA to provide a means for batch effect correction in the pre-processing of the data matrix:

- LIMMA's performs ANOVA to remove any measurable variation
- Two methods for unmeasurable batch effect correction with LIMMA

Principal Component Analysis (PCA)

- Perform singular value decomposition on the row-centered data
- Extract the first eigenvector and treat it as a covariate

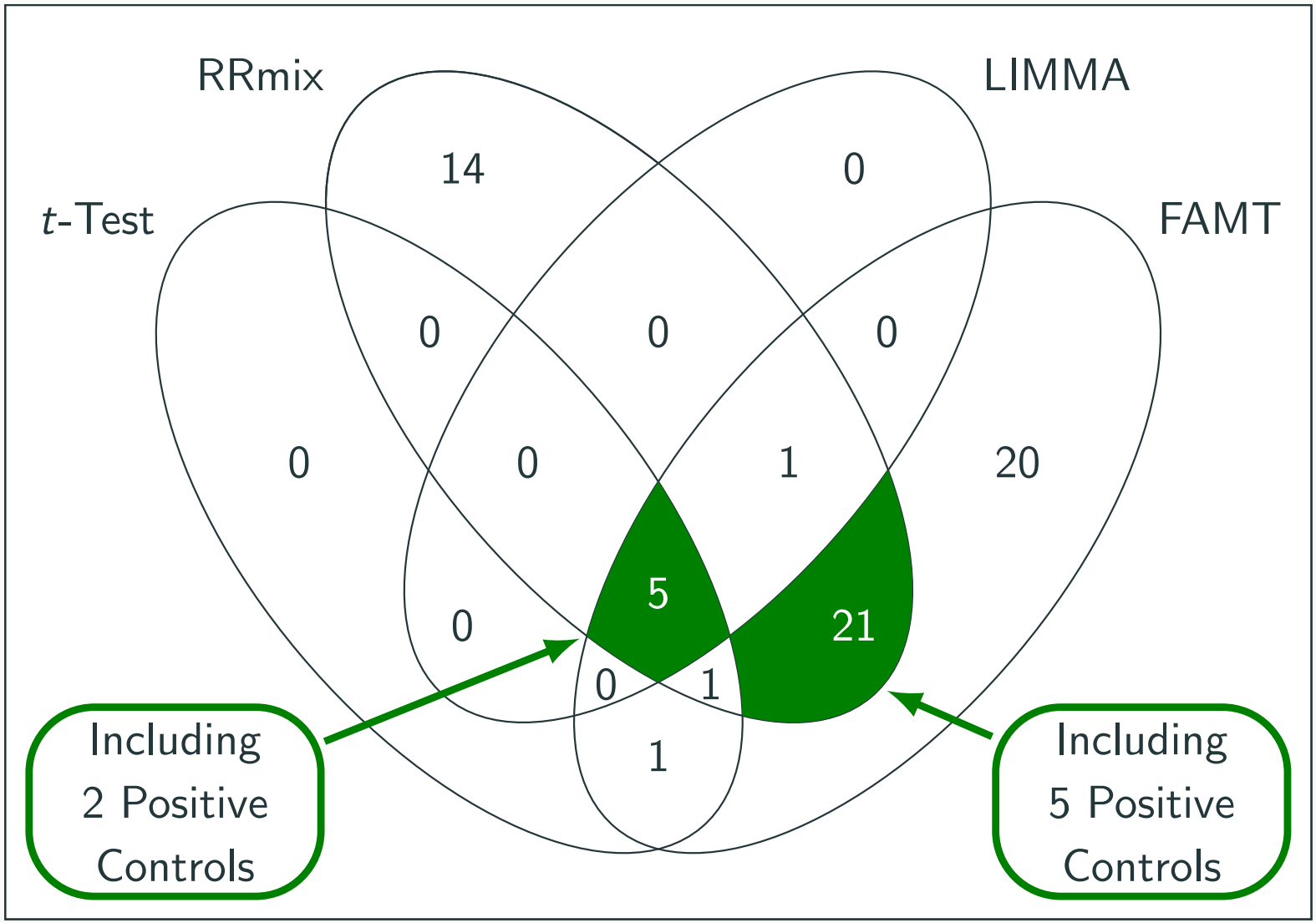
Surrogate Variable Analysis (SVA)

- Accounts for cross-compound dependencies induced by latent factors
- Data is modeled as a function of the predictor variable of interest
- SVD is then performed on the residuals to obtain eigenvectors
- Eigenvectors are tested for association with a significant proportion of the residual variation in the data
- Subset of metabolites associated with each significant eigenvector is determined
- Surrogate variables are calculated from this set of eigenvectors and the subset of the original data matrix for the differential metabolites

Factor Analysis for Multiple Testing (FAMT)

- Method in the family of latent factor models
- Does not include mixture component or prior assumptions on regression coefficients and compound-specific variances
  - Sparsity in the data is not modeled directly in FAMT
  - Accounted for by a post-hoc FDR thresholding step
- FAMT is a two-step procedure for model fitting and estimation:
  - Fitting is accomplished via the EM algorithm
  - Classification is done subsequently using approximate t-statistics

Bonferroni-Adjusted Significant Discoveries



Benjamini-Hochberg Adjusted Significant Discoveries

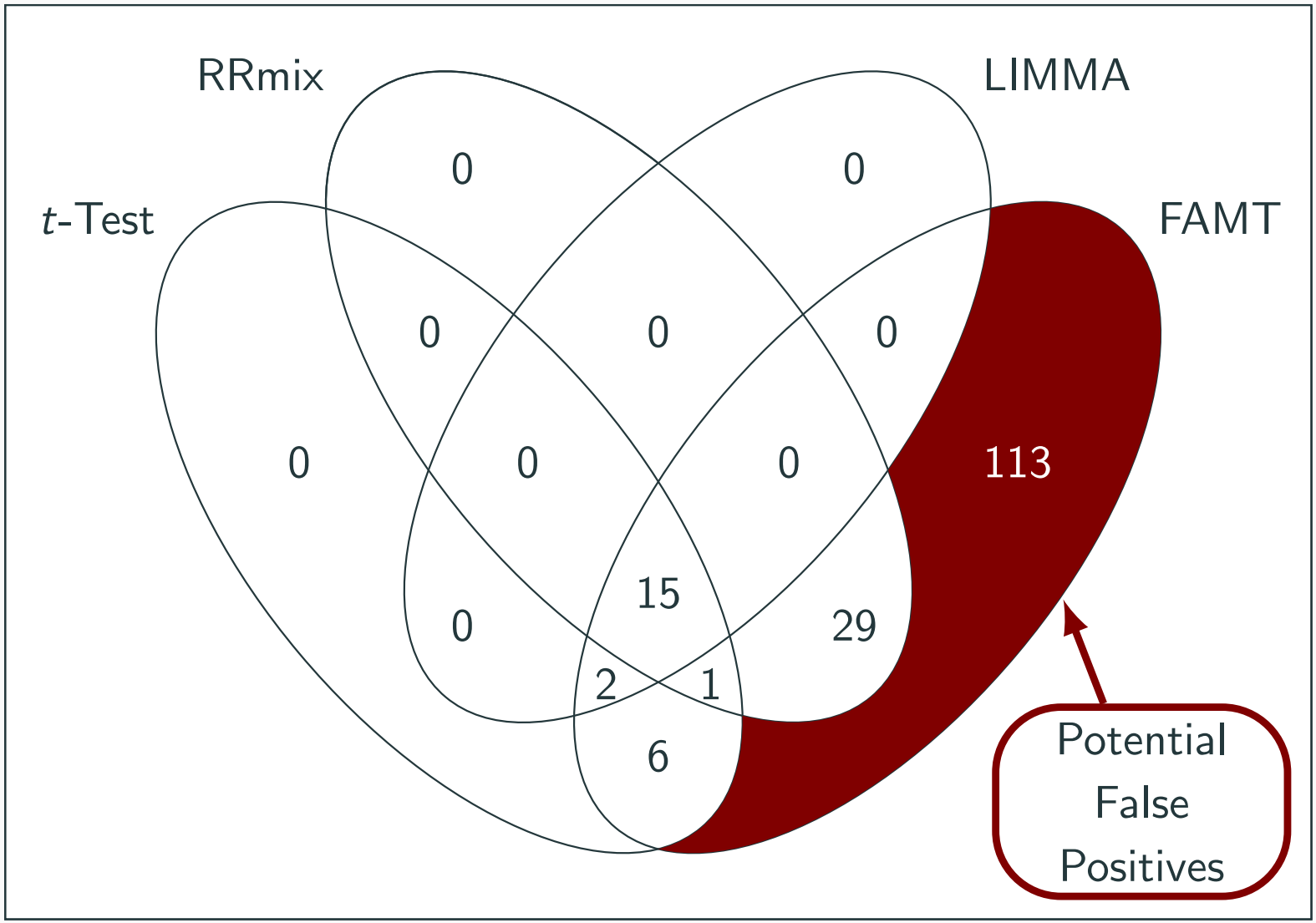


Table 1: Summary Results from Differential Abundance Analysis: Number of Total ($n = 265$) and Positive Control ($n = 8$) Discoveries

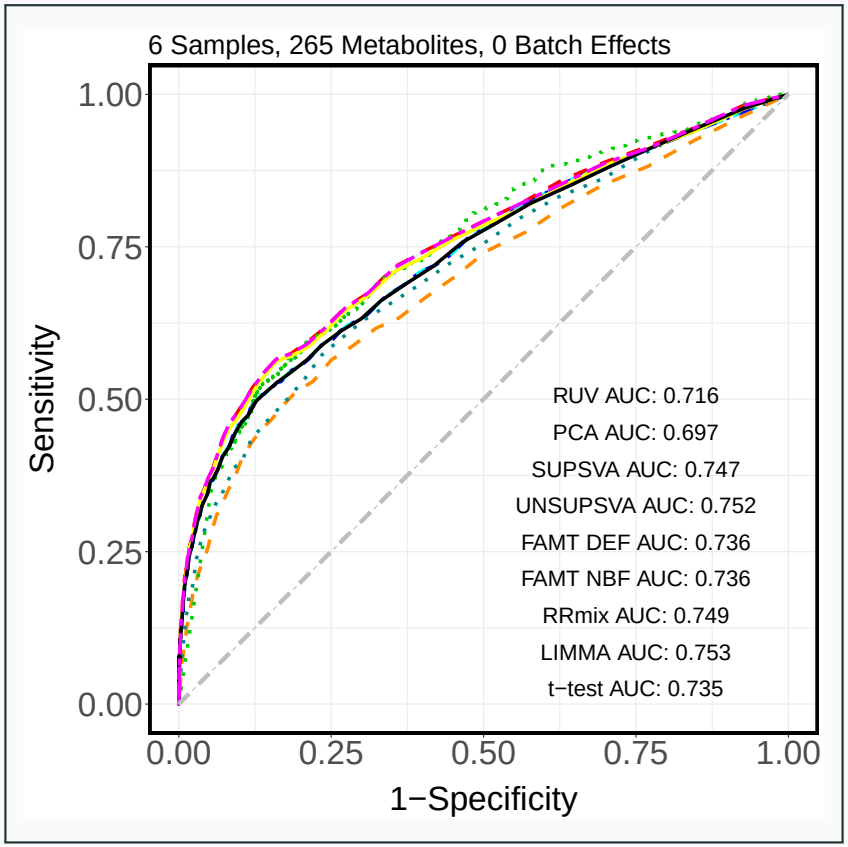
	No Batch Effects		With Batch Effects	
	Total	Controls	Total	Controls
t-Tests	49	5	24	2
LIMMA	115	6	19	4
PCA-LIMMA	158	5	119	7
SVA-LIMMA	152	7	114	7
FAMT	118	7	166	7
RRmix	39	6	42	7

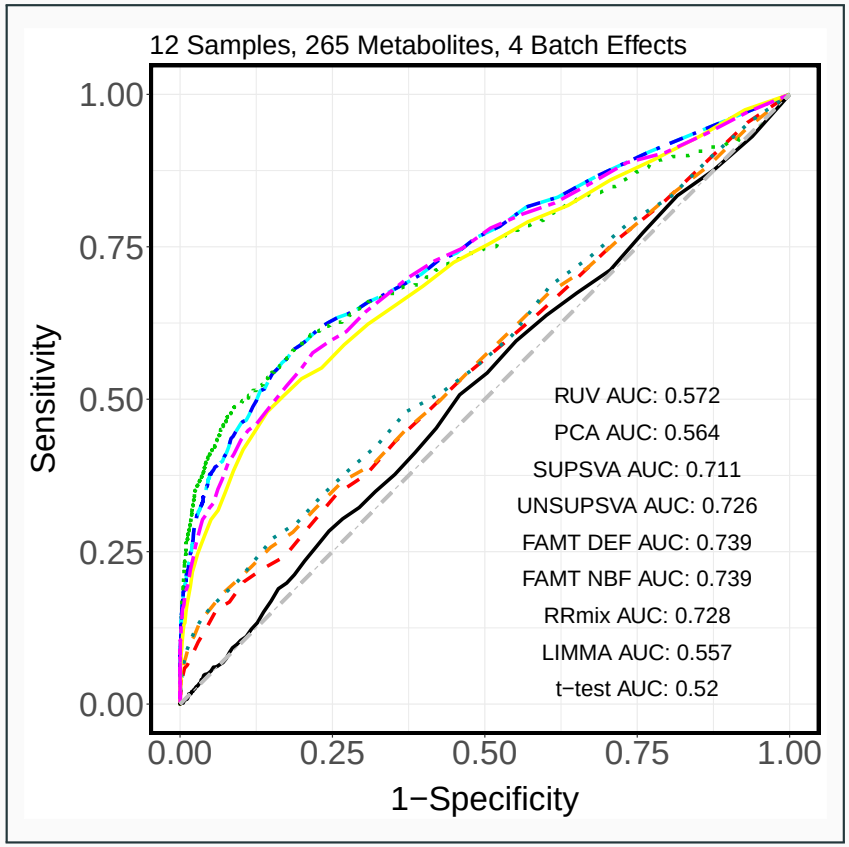
Simulation Study

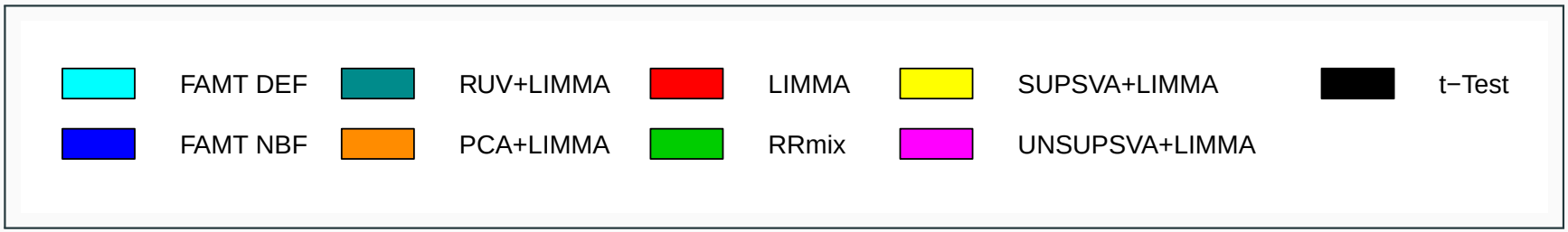
Conducted a series of four simulation studies using synthetic data, which closely mirror the LC-MS data set. For each study:

- Two sets of 50 simulated data sets were created:
  - One set with a sample size of $2n$ and four latent factors
  - One set with a sample size of n and no latent structure
- Data were simulated using estimates from original LC-MS data
  - 5% of metabolites had non-null status between treatment groups
  - 5% of metabolites simulated to resemble negative controls
- Sample size and number of metabolites increased in each study

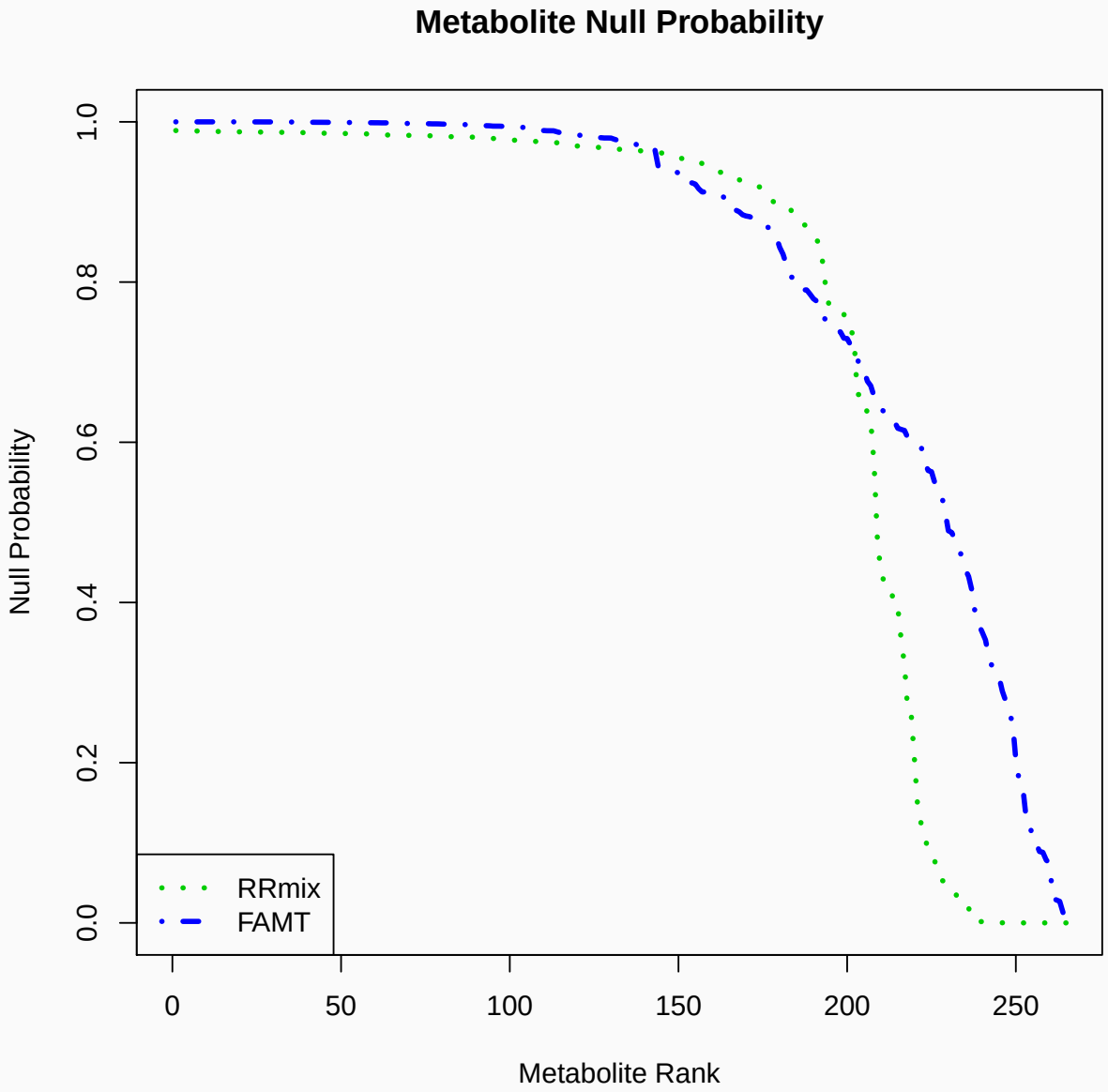
Average Receiver-Operator Characteristic (ROC) Curves







RRmix Thresholding



Conclusions

1. Operators induce major undesirable variation
2. Simple statistical methods are able to detect biological effects in the absence of batch effects
3. Latent factor models are able to detect biological effects in the presence of batch effects
4. Latent factor models facilitate combining of datasets for increasing statistical power
5. RRmix outperforms 2-stage procedures with respect to specificity

Applications to TB Metabolite Discovery: Methods

Prospective case control study to identify candidate urinary diagnostic biomarkers of active pulmonary TB:

- Participants enrolled at the GHESKIO center in Port-au-Prince Haiti
- Cases (110) matched to controls (102) by age, sex, and HIV status
- Clean-catch urine samples analyzed using LC-MS

Blinded validation cohort of 50 active pulmonary TB cases and 50 non-tuberculosis pulmonary disease controls analyzed for comparison

Applications to TB Metabolite Discovery: Results

Discovery (Haitian) Cohort:

- 49 metabolites significantly different in cases versus controls after False Discovery Rate correction
- 10 metabolites had AUC $> 85\%$ with 20×20 cross-validation
- MS/MS spectral analysis categorized 8 of the above metabolites

Validation (Vietnamese) Cohort:

- 4 of the 8 metabolites from discovery cohort had reliable sensitivity and specificity (AUC $> 75\%$)

Collaborators

Stephen Salerno, Jr.[1*], Mahya Mehrmohamadi[2,3*], Maria V. Liberti[3,4], Muting Wan[5], Martin T. Wells[5], James G. Booth[5], Jason W. Locasale[2,3,4]

1. Department of Biostatistics, University of Michigan, Ann Arbor, Michigan
2. Department of Molecular Biology and Genetics, Field of Genomics, Genetics and Development, Cornell University, Ithaca, New York
3. Duke Cancer Institute, Duke Molecular Physiology Institute and Department of Pharmacology and Cancer Biology, Duke University School of Medicine, Durham, North Carolina
4. Department of Molecular Biology and Genetics, Field of Biochemistry, Molecular, and Cell Biology, Cornell University, Ithaca, New York
5. Department of Statistical Science, Cornell University, Ithaca, New York

* These collaborators contributed equally to this work

Acknowledgements

- We thank Xiaojing Liu and Ahmad Cluntun for providing us with the experimental samples

- This work was supported by grants R00CA168997, R21 CA201963, R01CA193256 to J.W.L. and grant U19 AI111143 to M.T.W. from the National Institutes of Health (NIH)

- M.V.L was supported by predoctoral fellowships from the Sloan Foundation and the National Science Foundation

- M.M. was supported by T32GM007617 from the NIH and a Graduate Fellowship from the Duke University School of Medicine

Selected References

1. Gowda GA, Djukovic D. Overview of mass spectrometry-based metabolomics: opportunities and challenges. Methods in molecular biology. 2014.
2. Wehrens R, et al. Improved batch correction in untargeted MS-based metabolomics. Metabolomics. 2016.
3. Fernandez-Albert F, et al. Intensity drift removal in LC/MS metabolomics by common variance compensation. Bioinformatics. 2014.
4. Mehrmohamadi M, Liu X, Shestov AA, Locasale JW. Characterization of the usage of the serine metabolic network in human cancer. Cell reports. 2014.
5. De Livera AM, Sysi-Aho M, Jacob L, et al. Statistical methods for handling unwanted variation in metabolomics data. Analytical chemistry. 2015.
6. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology. 2004.
7. Wan M. Model-Based Classification with Applications to High-Dimensional Data in Bioinformatics. PhD Dissertation,. 2015; Field of Statistics, Cornell University.
8. Causeur D, et al. Factor analysis for multiple testing: an r package for large-scale significance testing under dependence. Journal of statistical software. 2011.
9. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics. 2007.

And now for something completely different

Translating ESRD Patients' Quality of Care:

Dialysis Facility Compare and the 5-Star Rating System

Stephen Salerno

Department of Biostatistics · University of Michigan

March 14, 2018

Chronic Kidney Disease

Chronic Kidney Disease (CKD) is a severe condition afflicting over 30 million US adults. CKD is characterized by:

- A failure of the kidneys to properly filter waste from the blood
- Deteriorating function leading to End Stage Renal Disease (ESRD)

Transplantation is preferred but often not possible due to shortages in donor organs and a necessity to be a viable match

- Dialysis performs the functions of the kidney through an apparatus
- Dialysis has universal healthcare coverage in the U.S. today

Dialysis: A Life-Sustaining Therapy

Dialysis uses a chemical solution to remove wastes, salt and extra water from the blood. There are two main types:

- Hemodialysis (HD):
  - Uses an external apparatus (dialyzer) to filter blood
  - Three common forms: in-center, in-center nocturnal, home-hemo
- Peritoneal Dialysis (PD):
  - Uses lining of abdominal cavity (peritoneal membrane) to filter blood

Dialysis facilities are required to report several metric related to the dialysis adequacy, modality, vascular access, and other patient health outcomes

Brief Timeline of Key Events

October 1972	Medicare ESRD Program
August 1997	Balanced Budget Act
January 2001	Dialysis Facility Compare (DFC) Site
March 2010	Patient Protection and Affordable Care Act
January 2015	Original 5-Star Rating System
April 2015	Star Rating Technical Expert Panel (TEP)
October 2016	Updated 5-Star Rating System
February 2017	Star Rating Technical Expert Panel (TEP)
October 2018	Second Update to the 5-Star Rating System

University of Michigan Kidney Epidemiology and Cost Center

The Kidney Epidemiology and Cost Center (KECC) is a major research center within the University Of Michigan School of Public Health with:

- Epidemiological, clinical, public policy, and economic research relating to ESRD, CKD, and organ transplantation
- Funding through multiple government and private sources, including Centers for Medicare and Medicaid Services (CMS)
- Data on $\sim$2.5 million ESRD patients drawn primarily from Medicare

After joining in August 2016, I worked on developing and implementing two updates to the Star Rating methodology, in addition to other research for the End Stage Renal Disease clinical quality measure development team

DFC 5-Star Rating System

KECC developed a 5-Star Rating System in 2014:

- Utilizing clinical quality measures reported on the DFC website
- To rate the quality of care provided by dialysis facilities
- To provide patients, families, and caregivers information to easily compare dialysis facilities

The 5-Star Rating System is currently in its third iteration of policy-based and methodological updates

Original DFC 5-Star Rating System

Raw measures differ in distribution & scale

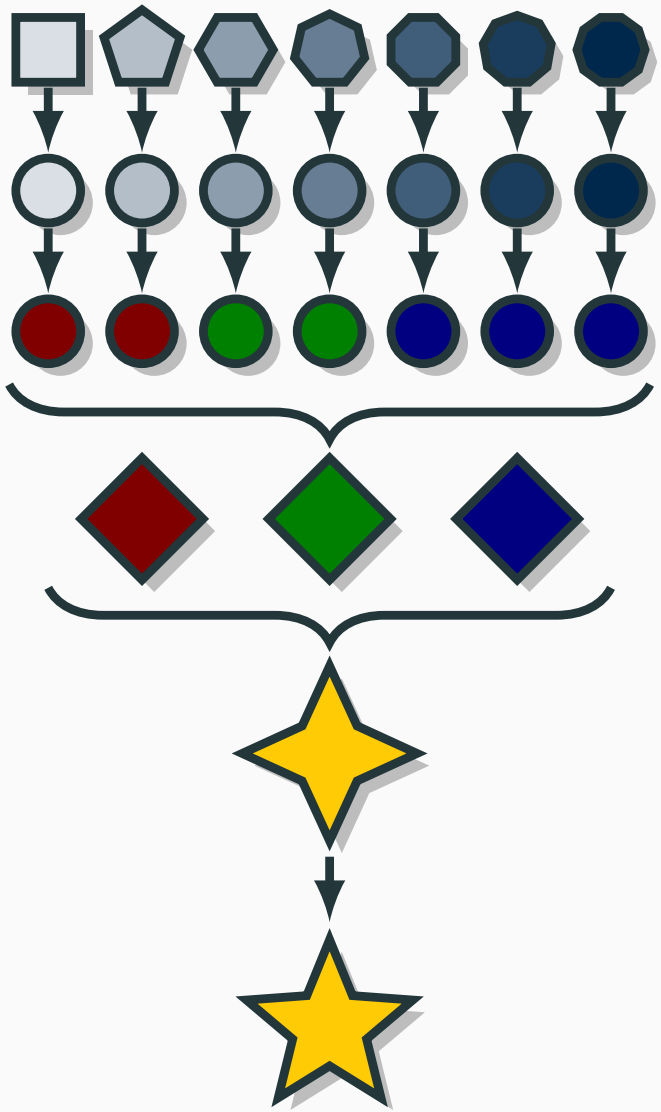
Measures are transformed via probit scoring

Factor analysis identified 3 measure domains

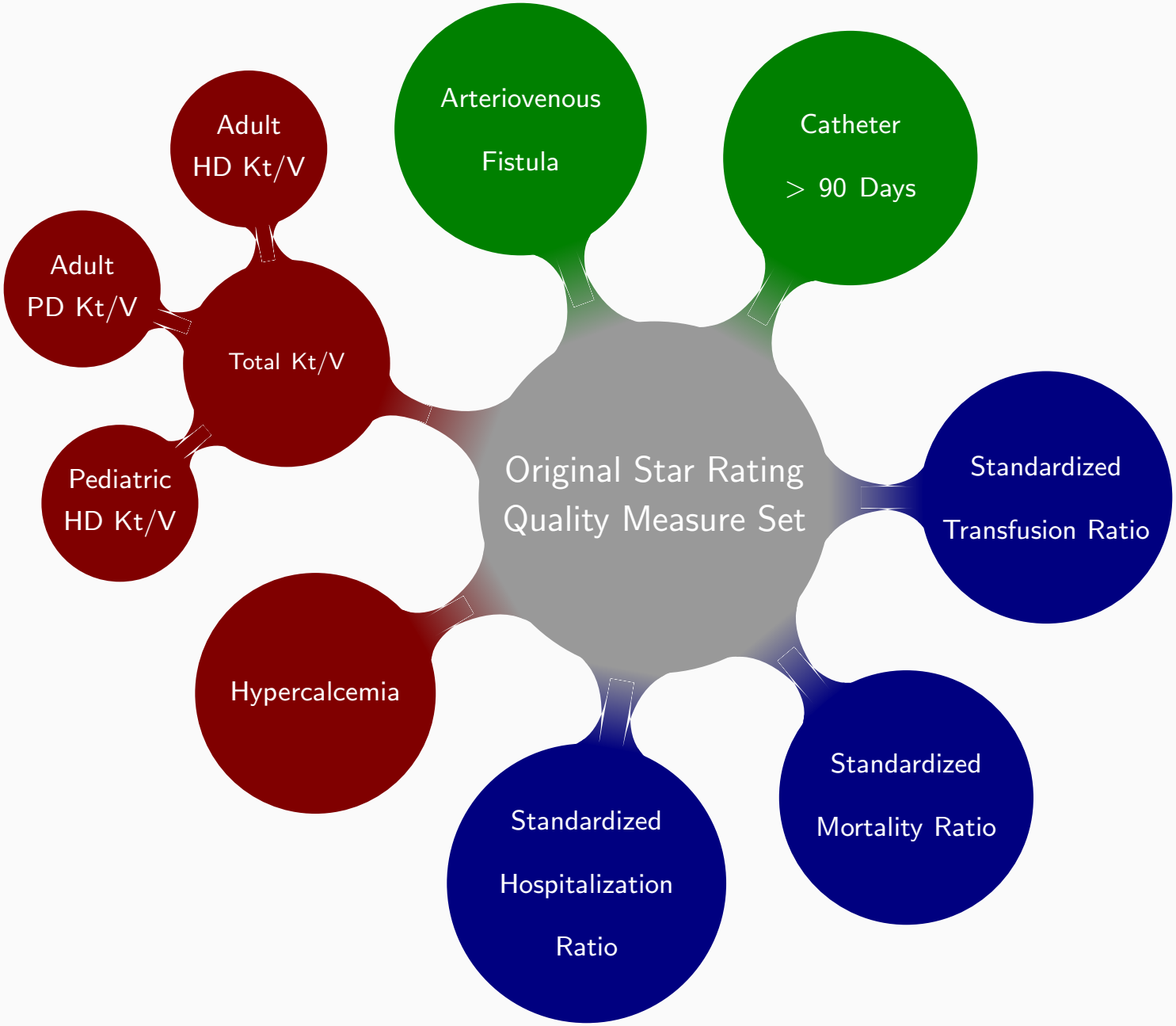
Measures within domains are averaged

Domain scores are averaged into a final score

Final scores are grouped into Star Ratings



Original Quality of Patient Care Clinical Measure Set



Quality of Care Star Rating Assignment


Much Above Average (10%)


Above Average (20%)


Average (40%)


Below Average (20%)


Much Below Average (10%)

Technical Expert Panel Recommendations and Updates (2015)

Establish a Baseline to Show Improvement

- Account for changes in facility performance over time
- Compare data to performance standards set in a baseline year

Account for Highly Skewed Measures

- Limit impact of extreme scores
- Ensure star ratings are not determined by a single measure

Keep the Continuity of Measures

- Ensure the accuracy of the ratings

Establish a Baseline to Show Improvement

Baseline Year

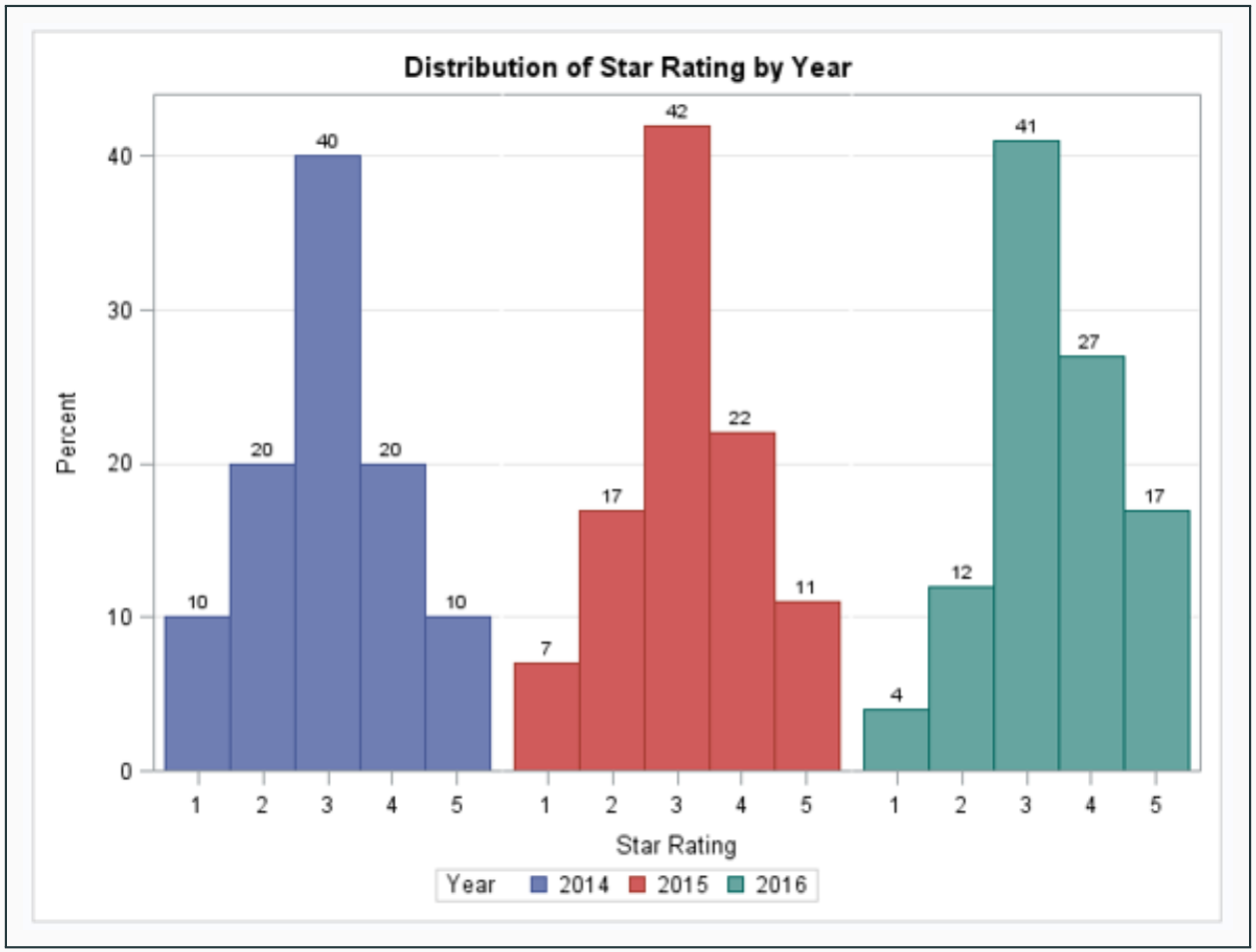
- Final scores in a baseline year used to determine final score cutoffs
- Baseline year cutoffs set based on 10%, 20%, 40%, 20%, 10% rule

Current Year

- Baseline year cutoffs used to assign ratings in a current year
- Defining baseline cutoffs allows for observing performance over time

Baseline Scoring Implications

Absolute measure value improvement guarantees improvement in facility scores, but compress the Star Ratings and lessen discriminatory power



Percentage Measures

- Truncated Z-scores calculated for Kt/V, Hypercalcemia, Fistula, & Catheter first in baseline year, then applied to current year
- Winsorization performed so final measure scores have mean = 0, SD = 1, range = [-2.58, 2.58]
- Limited range prevents Star Ratings from being determined by outlier performance on a single measure
- Variance stabilization ensures measures influence rating equally

Standardized Ratio Measures

- SMR, SHR, and STrR are multiplied by an adjustment factor to:
  - Account for differences in population event rates between years
  - Allow current year ratio values to reflect same measure values that would have been observed in the baseline year
- Probit transform applied in baseline year – defines criteria that assigns scores to measure values
- This criteria is then applied in the current year for reporting, after implementation of the adjustment factor

Technical Expert Panel Recommendations and Updates (2017)

Differentiate Rebaselining & Resetting the Star Rating Distribution

Rebaselining: Rescoring of measures when establishing a new baseline

Resetting: Determining new cutoffs for the entire Star Rating distribution

Define Criteria for Rebaselining

- When new measures are added or removed
- When current measures are updated

Define Criteria for Resetting

- When the Star Rating distribution is significantly compressed
- When the information provided by individual measures is no longer useful in discriminating facility-level performance

Measure Inclusion and Rebaselining

- Adding or removing measures changes the set of clinical quality features in which facilities are compared
- Addition of previously unmeasured features prevents year-to-year comparison before the first year of collection
- A new measure set might follow a different domain construct after updated factor analysis
- Quality of care standards may differ between measures

Define Criteria for Rebaselining

- Update the measure Set at predictable time intervals
    - Initially estimated 3-year intervals
- Evaluate Star Rating & individual measure distributions
- Include important updates & novel measures immediately

Expanded Measure Set Considerations for October 2018

New Measures Proposed for Addition

- Pediatric PD Kt/V
- Standardized Readmission Ratio (SRR)
- In-Center Hemodialysis Consumer Assessment of Healthcare Providers and Systems (ICH CAHPS) Measures

Current Measures with Updated Definitions

- Standardized Fistula Rate (Updates Current Fistula)
- Long-Term Catheter Rate (Updates Current Catheter $>$ 90 Days)
- Standardized Transfusion Ratio (STrR)
- Standardized Mortality Ratio (SMR)
- Standardized Hospitalization Ratio (SHR)

Patient Reported Outcomes Star Rating

Summary of ICH CAHPS Data

- 6 measures: 3 global & 3 composite derived from survey questions
- Analysis revealed strong correlation within CAHPS measures, but lack of correlation with clinical measures
- High levels of missing data at the facility level ($\sim$50%)

Recommendations Regarding ICH CAHPS Inclusion

- Recommended ICH CAHPS receives separate Star Rating on DFC
- ICH CAHPS measures will be transformed into linearized scores

Defining Criteria for Resetting: Active Research

Current area of analysis to inform policy decisions:

- Analyze the Star Rating distribution for significant compression
- Determine if distributional movement is due to location (mean) shift or skew (compression)
  - Shift suggests facility performance is out-pacing baseline standards
  - Skew suggests individual measures are "topped-out" and cannot discriminate facility performance

Consider resetting the Star Rating distribution when the number of 1- and 2-star facilities falls below a certain proportion

Summary of Star Rating Updates

2015 Original Star Rating

First provided patients a means to compare dialysis facilities based on an overall summary of several key clinical quality measures

2016 Star Rating Methodology Update

Better aligned the DFC Star Rating System to the needs and preferences of patients, patient advocates, providers, and other stakeholders

2018 Star Rating Methodology Update

Will expand the clinical quality measure set & fine-tune the 2016 methodology update, addressing key methodological & policy issues as the system continues to develop

Other Current Areas of Research

- How to handle missing facility-level clinical measure values?
- Should care domains be weighted by a metric of importance?
- What is driving facility improvement over time?
- Does facility performance differ by geographical region, provider type, facility size, ...?
- Are there better ways to differentiate facility performance?
- What other clinical aspects (if any) should be included in ratings?

Selected References

1. Centers for Disease Control and Prevention. Chronic kidney disease surveillance system website. http://www.cdc.gov/ckd.
2. Chen, J. Public reporting improves healthcare. Health Information, 2010.
3. Pozniak, A., Pearson J. The Dialysis Facility Compare Five-Star Rating System at 2 Years. Clinical Journal of the American Society of Nephrology. 2017 Dec 18.
4. End Stage Renal Disease Dialysis Facility Compare Star Ratings Technical Expert Panel: Summary Report. (2015, April 27-28). In-Person Meeting. https://dialysisdata.org/content/esrd-measures
5. End Stage Renal Disease Dialysis Facility Compare Star Ratings Technical Expert Panel: Summary Report. (2017, February 21). In-Person Meeting. Retrieved from https://dialysisdata.org/content/esrd-measures
6. Kidney Epidemiology and Cost Center. Technical Notes on Dialysis Facility Compare Star Rating System. Retrieved from https://dialysisdata.org/content/methodology
7. Kidney Epidemiology and Cost Center. Technical Notes on the Updated Dialysis Facility Compare Star Rating Methodology. Retrieved from https://dialysisdata.org/content/methodology
8. Kidney Epidemiology and Cost Center. Technical Notes on the Dialysis Facility Compare Star Rating Methodology for the October 2018 Release. Retrieved from https://dialysisdata.org/content/methodology

Thank You and Happy π Day!

Supplementary Appendix

Current Urinary Diagnostic Biomarkers of TB

Urine tests are an increasingly common modality used to enable non-invasive, rapid and point-of-care diagnosis of various infectious diseases:

- Lipoarabinomannan (LAM) is a component of the mycobacterial cell wall that is shed into urine and capable of being detected in the urine of patients with active pulmonary TB
- Overall estimated sensitivity of 46% and specificity at 89%
- Test has been associated with a decrease in all-cause mortality in HIV and Mtb co-infected participants with a relative risk reduction of 17%

Current KECC Projects

- *Quality Measure Development, Maintenance, and Support*

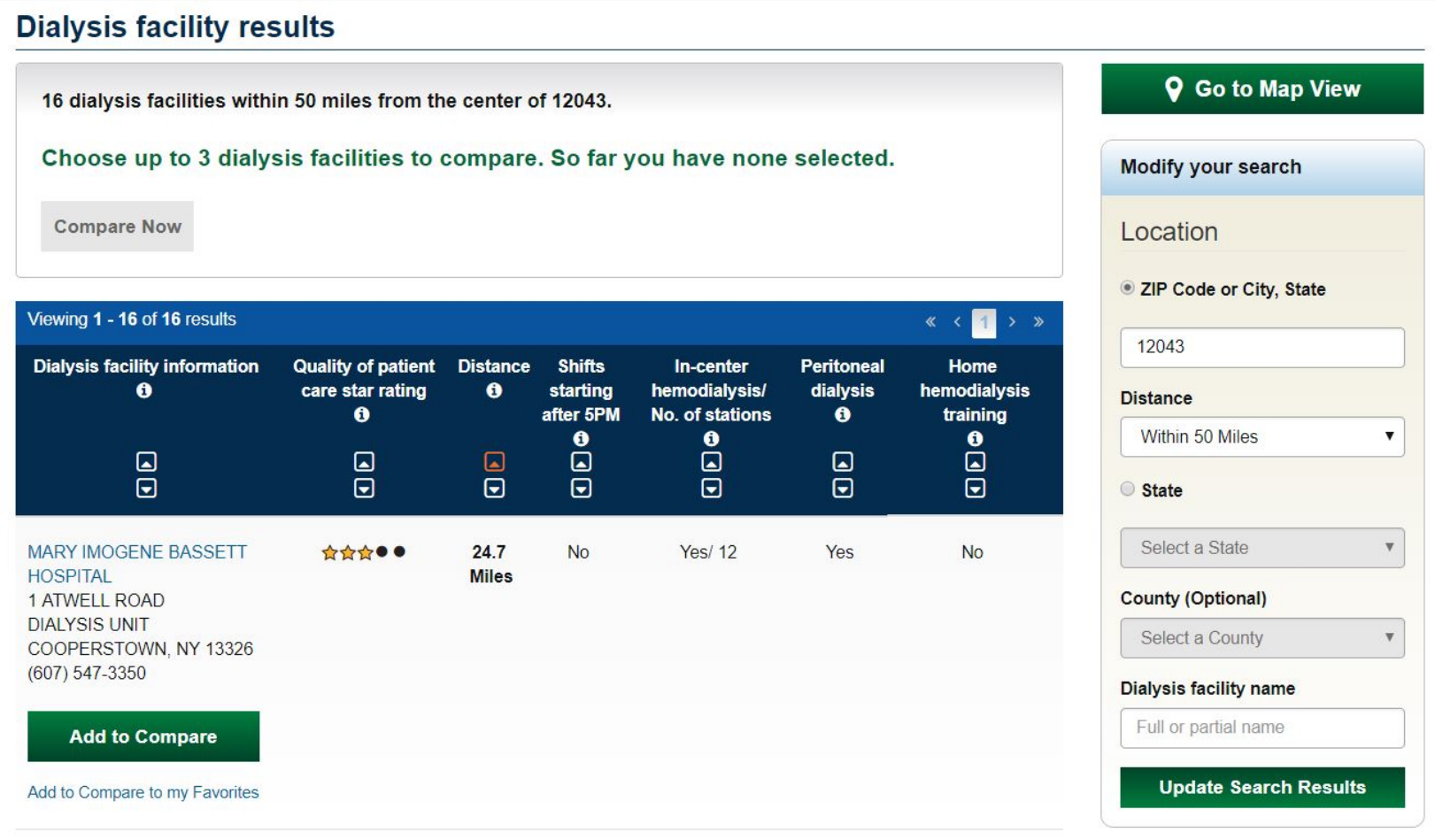
- Utilization of Data Indicators in the ESRD Survey Process
- Evaluation of the Comprehensive ESRD Care (CEC) Initiative
- End Stage Renal Disease (ESRD) Quality Incentive Program (QIP)
- United States Renal Data System Coordinating Center (USRDS)
- ESRD, Data, Registries, Acute kidney injury, Chronic Kidney Disease, dialysis, Kidney transplant Optimization and Simulation of Kidney Paired Donation Programs (Paired Donation)
- Cascading Impact of Bundled Payment on Peritoneal Dialysis Provisions and Outcomes Research Institute
- Enhancing the cardiovascular saftey of hemodialysis care: Patient-Centered Outcomes Research Institute
- Supporting, Maintaining and Improving the Surveillance System for Chronic Kidney Disease in the U.S.

Dialysis Facility Compare

Figure 1: The Medicare Dialysis Facility Compare Site

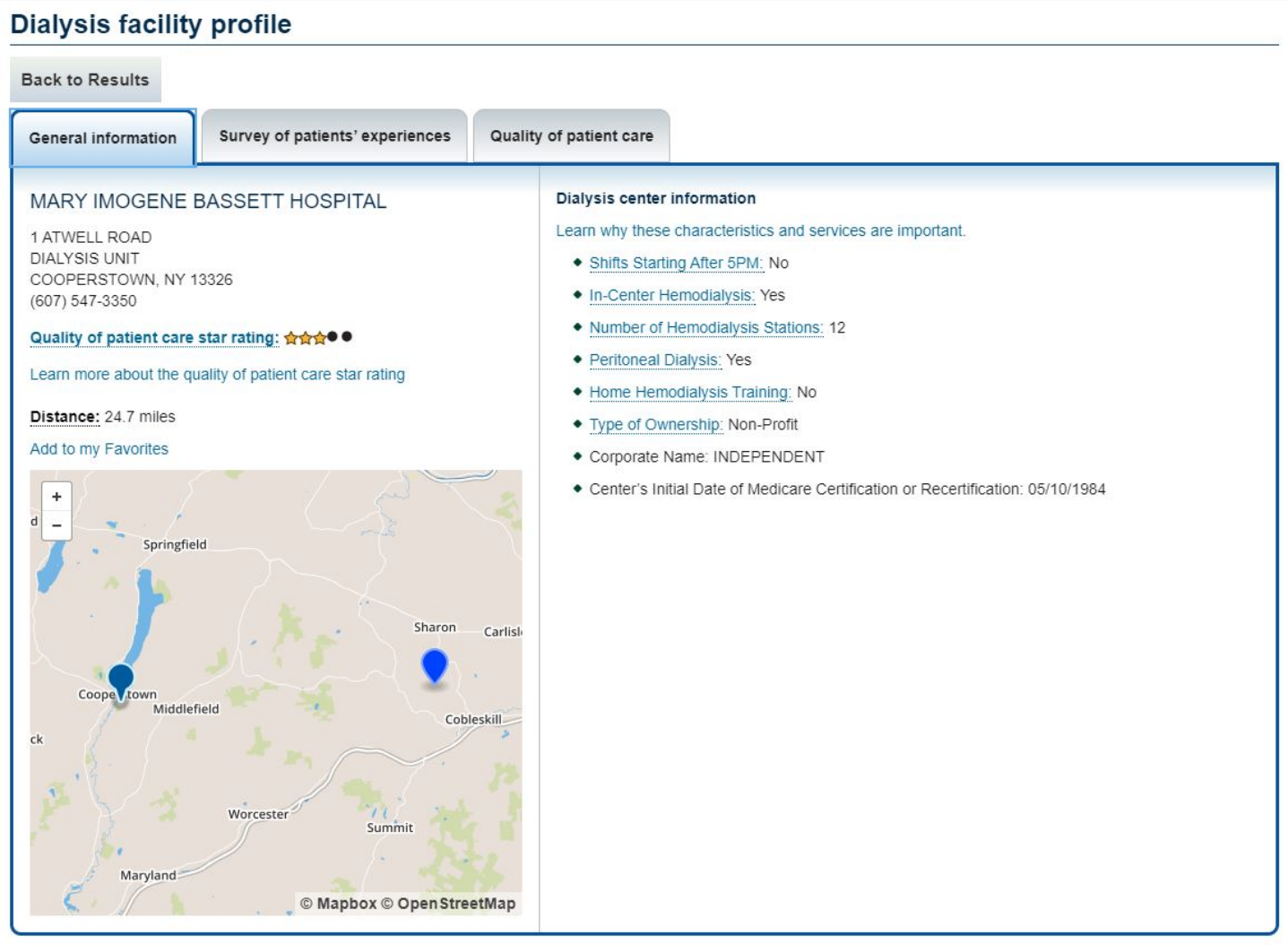
Example Search

Figure 2: Search Results for My Hometown Zip Code



Example Facility Profile

Figure 3: Profile of Nearest Facility to My Hometown



Example Quality Measure Report

Figure 4: 5-Star Rating and Clinical Quality of Patient Care