Phenotyping the Population-based Administrative Databases: Applications in Cancer, Heart Failure and Frailty

MERCK

Team

- Richard Baumgartner, Yue-Ming Chen, Vladimir Svetnik (BARDS, Biometrics Research)
- Jinghua He, Mehmet Burcu, Sarah Liu, Ed Bortnitchak (CORE, Pharmacoepidemiology)
- Weifeng Xu, Ben Koch, Margaret Coughlin (BARDS, Statistical Programming)

MERCK

Outline

- Increasing importance of RWD/RWE in early and late drug development
- Diverse sources of the RWD used to generate RWE
- Phenotyping of Electronic Medical Records (EMRs) / Electronic Health Records (EHRs) and medical claims – background, workflow and challenges
- Examples of phenotyping
  - Frailty
  - Heart Failure
  - Cancer stage (as a proof of concept)
- Challenges, Future Developments and Broader View of AI in enrichment of RWD

MERCK

Explanations of RWD/RWE

- **Real-World Data (RWD)** are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

- **Real-World Evidence (RWE)** is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD.

MERCK

Changing Landscape from Early to Late Stage Development Towards Acceptance of RWD/RWE

"Real-world evidence is defined as the analysis of RWD in a study designed with a high degree of pragmatism, regardless of study type. A wide variety of study designs can be used to generate this evidence, including pragmatic RCTs. Many questions about a drug remain unanswered at the time of approval; some of them involve optimal dosing regimen, longer-term outcomes, and outcomes in various subpopulations. It is not feasible to answer all of these questions with traditional RCTs. Using RWE to begin to address these questions is preferable to having no evidence whatsoever."

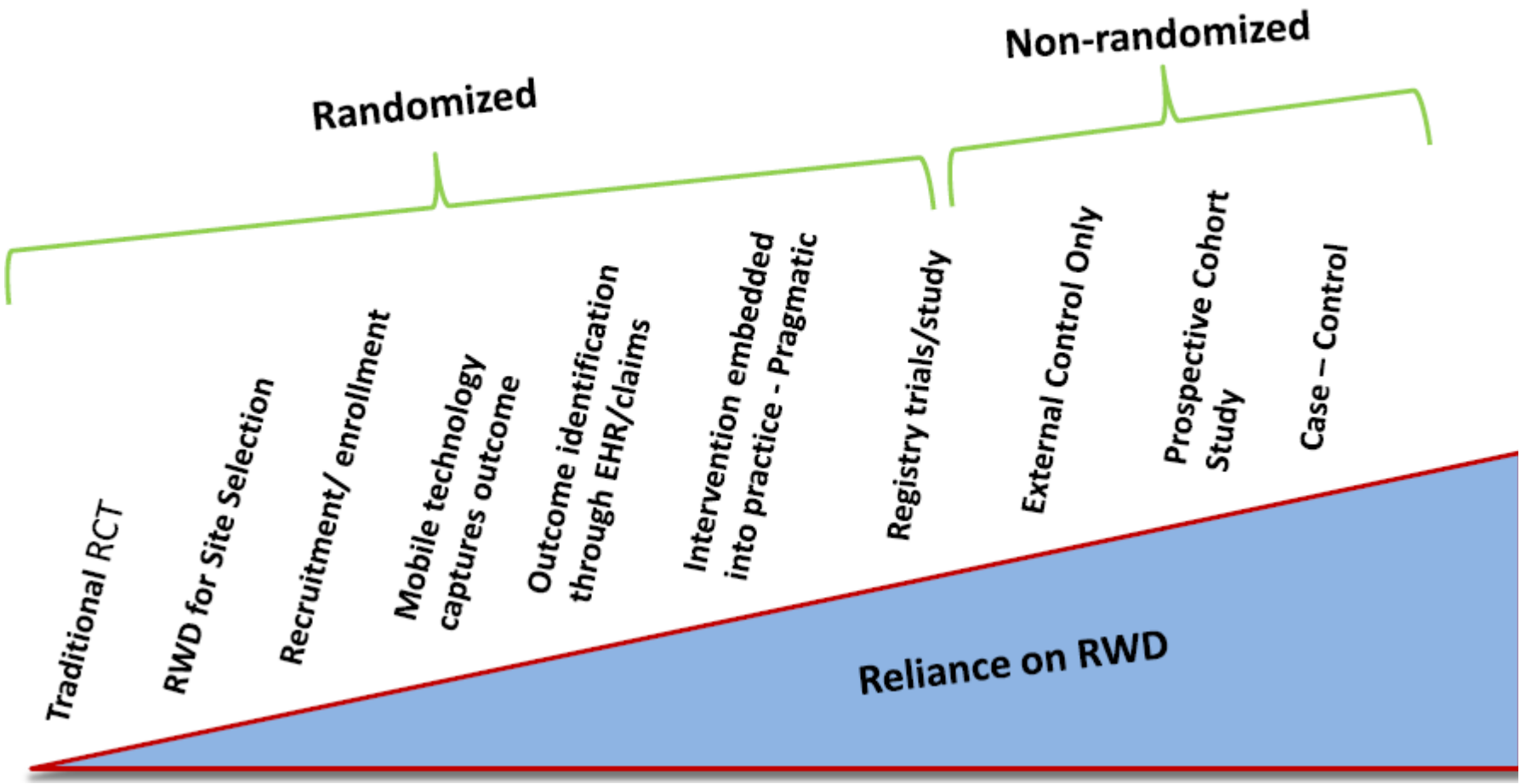
Jarow JP, LaVange L, Woodcock J: Multidimensional Evidence Generation and FDA Decision Making, JAMA, 2017

"… there is nothing in our statute or regulations that prevent FDA from using a broad range of informative sources of evidence. On the contrary many of our statutory responsibilities boil down to one principal calculus, what we do know and how we balance benefit and risks on the fullest possible information."

Scott Gottlieb, FDA commissioner:
http://nationalacademies.org/hmd/Activities/Research/DrugForum/2017-SEP-19.aspx

MERCK

Spectrum of Reliance on RWD



Jacqueline Corrigan-Curay, Real world evidence and path forward FDA, Duke Margolis conference, 2017

MERCK

Timeline of Increasing Importance of the RWD/RWE

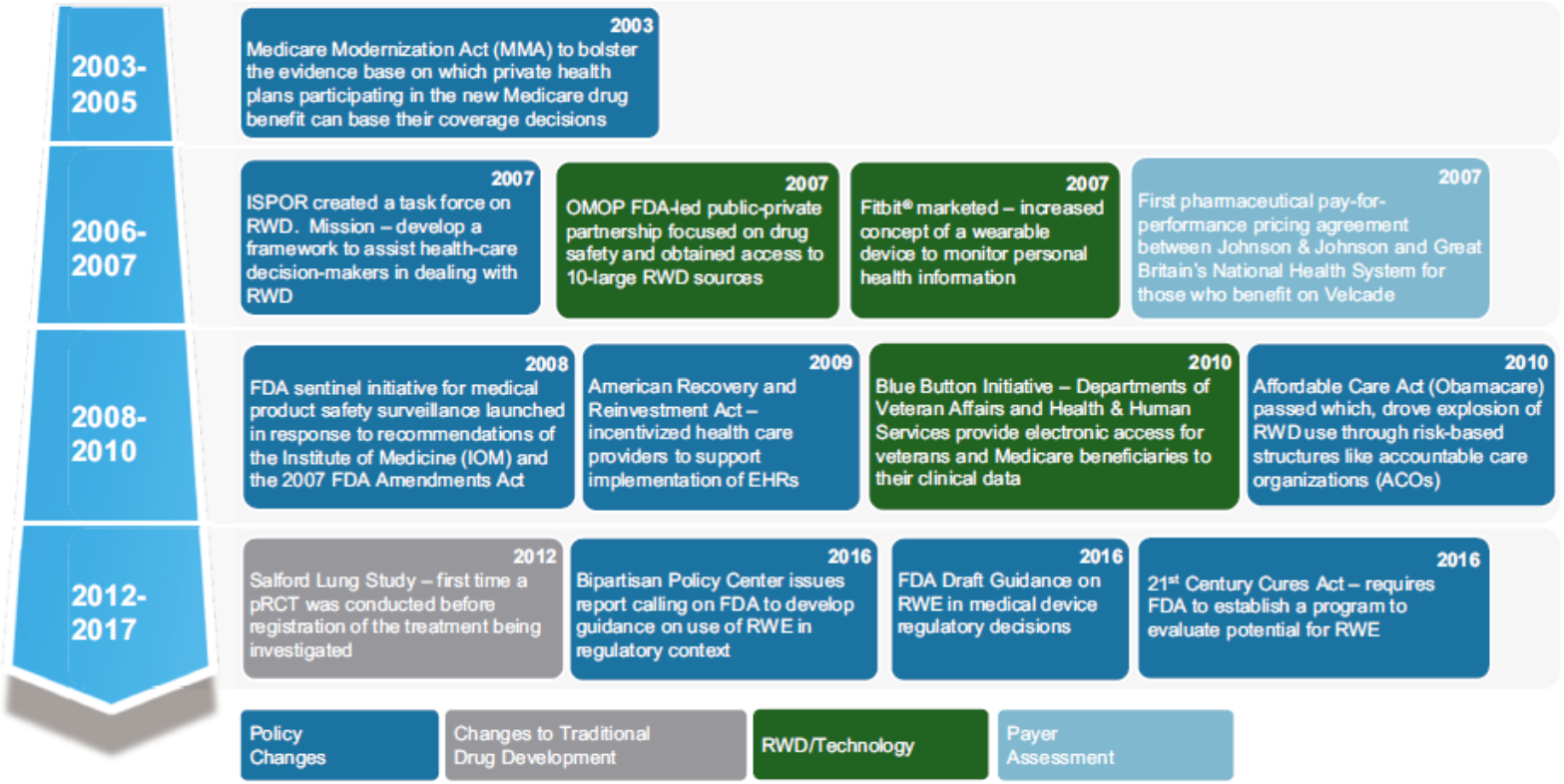


Figure 2 Key events leading to the importance of real-world evidence. International Society for Pharmacoeconomics and Outcomes Research (ISPOR); Observational Medical Outcomes Partnership (OMOP); pragmatic randomized clinical trial (pRCT); real-world evidence (RWE); real-world data (RWD); electronic health record (EHR).

Swift et al. Clin. Transl. Sci., 11, 450-60, 2018

MERCK

Sources of RWD

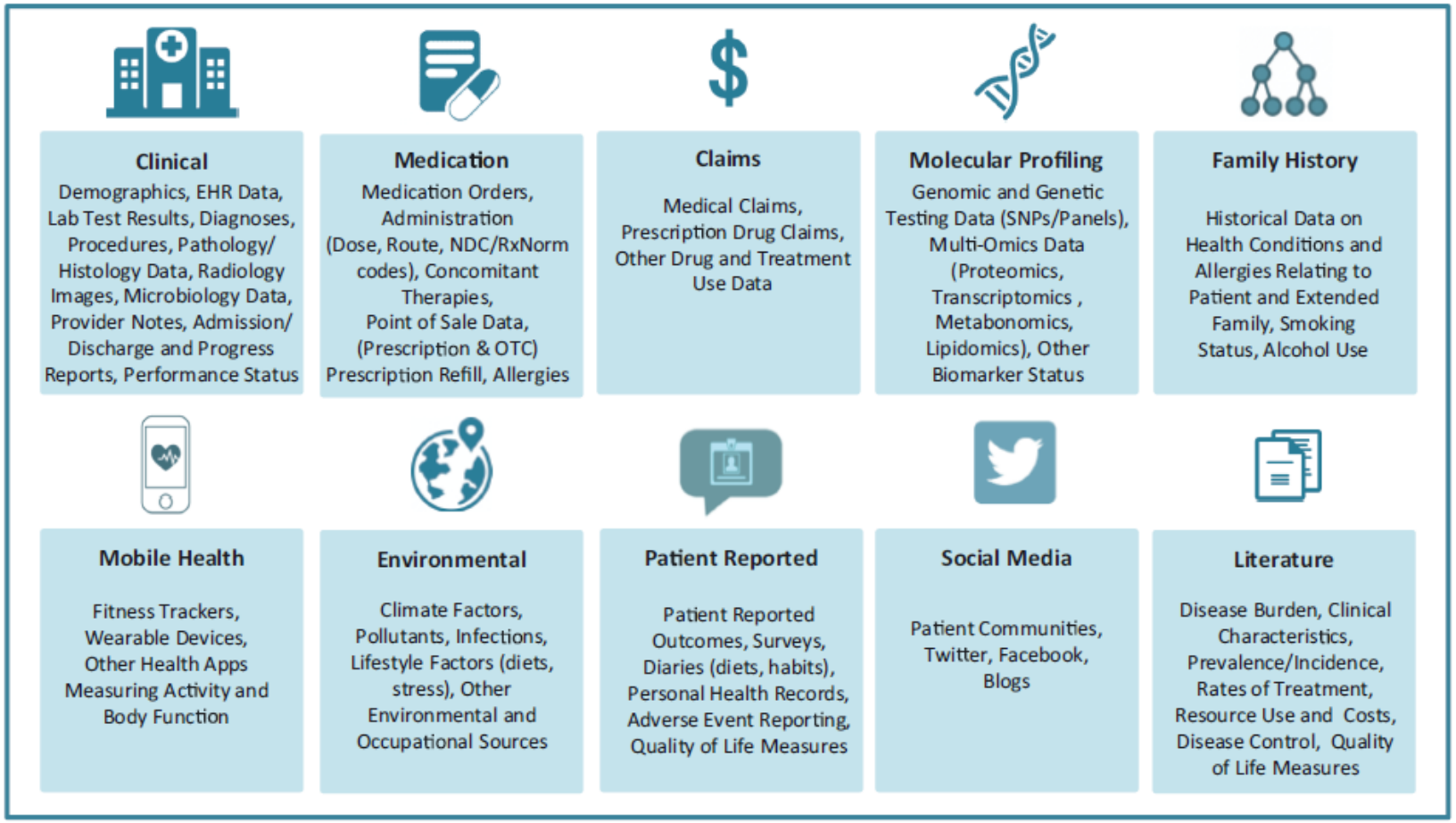


Figure 1 Types of real-world data.

Swift et al. Clin. Transl. Sci., 11, 450-60, 2018

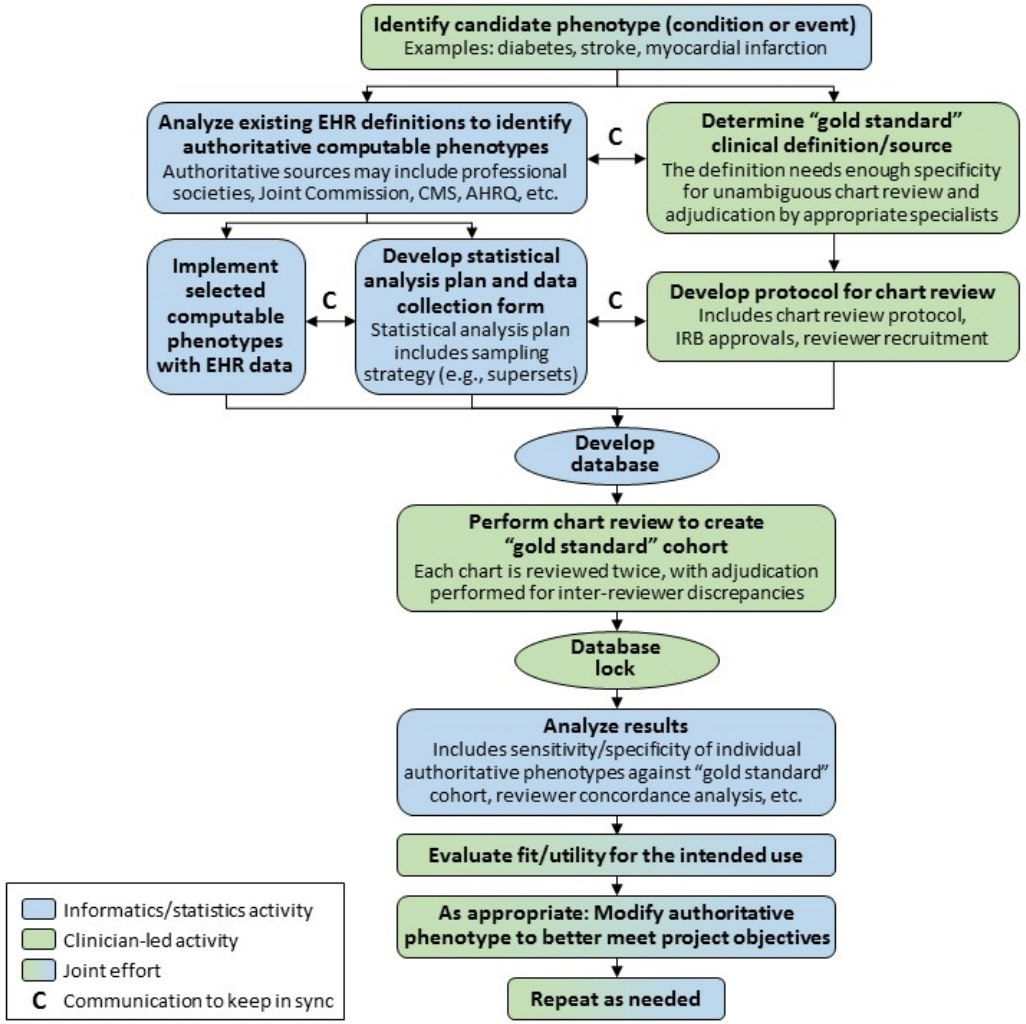
MERCK

Phenotyping of RWD
Concept and Challenges

- Frequently only partial information on important patient descriptors (phenotypes) is available in the RWD (EMR/EHR and medical claims)
- Examples of RWD phenotypes may include cancer stage, frailty, heart failure and osteoporosis diagnosis, diabetes mellitus type 2 status, etc.
- Algorithms to ascertain phenotypes developed from RWD databases may be based on prior knowledge such as clinical decision trees
- Recently, machine learning techniques have been applied in this context with great promise
- Challenges:
  - Relevant or in some cases indispensable confounders are not present or poorly recorded in RW data bases
  - Availability of the linked data-bases (gold standard) to derive desired phenotype

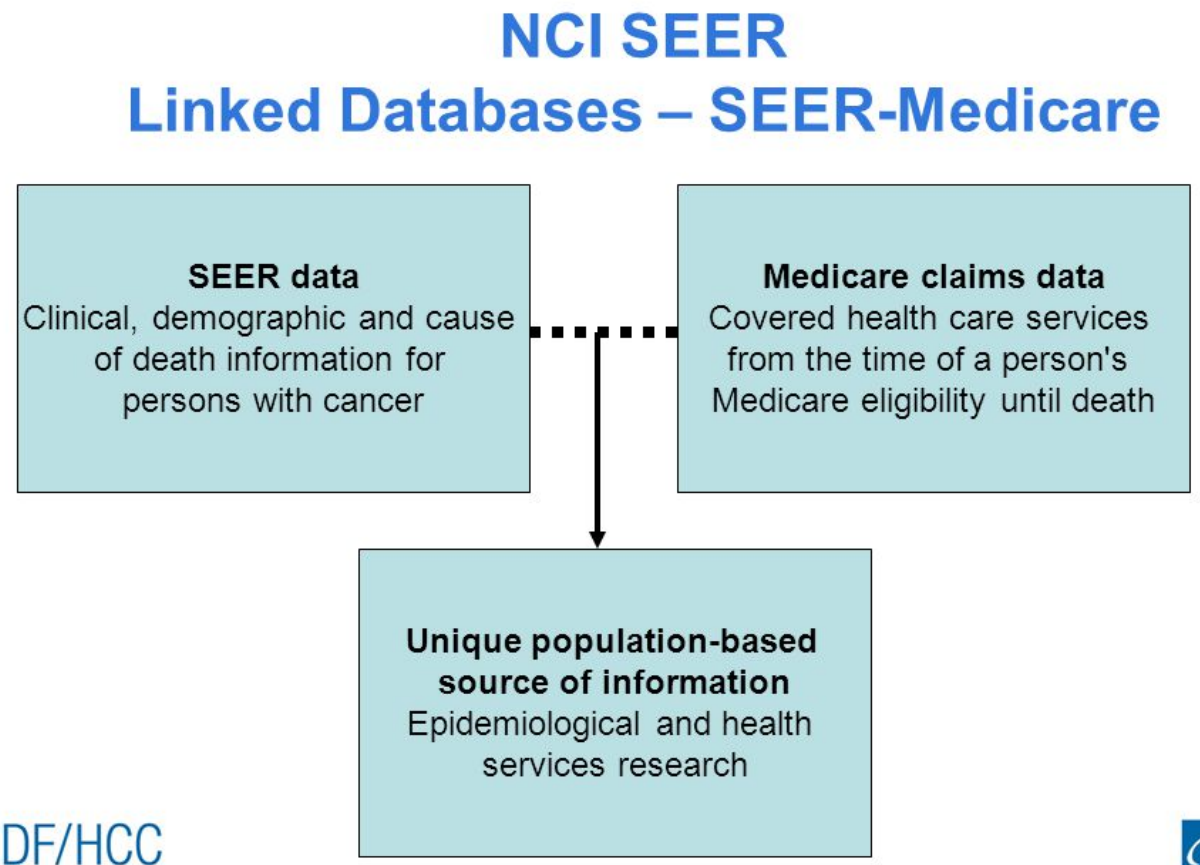
MERCK

Workflow of Phenotyping RWD on an Example of EHRs



https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/

Linked Data Bases – Key Requirement to Successful Phenotyping



DF/HCC
DANA-FARBER/HARVARD CANCER CENTER
A Comprehensive Cancer Center Designated by the National Cancer Institute

UMASS BOSTON

Carol Lowenstein, Cancer registries and medical records rich data sources. https://slideplayer.com/slide/5155954/

MERCK

Example 1:
Phenotyping of Frailty

- Frailty – physiological state of increased vulnerability to resulting from age and disease decreased physiologic reserves in multiple systems
- Unmeasured or residual confounding by frailty is an open and potentially profound challenge in analysis of administrative claims data
- Frailty hallmarks
  - Nutritional decline, fatigue, decrease activity and weakness
- Frailty vs. comorbidity – frailty is not linked to specific disease codes
- Frailty is strongly associated with the adverse effect (AE) of health outcomes including mortality
- Operational definition of frailty – based on dependency in activities of daily living (ADL) thus representing a suitable proxy for frailty

MERCK

Example 2:
Phenotyping of Heart Failure from EHR/Medical Claims Data

- HFrEF vs HFpEF
  - Heart Failure with Reduced Ejection Fraction vs. Heart Failure with Preserved Ejection Fraction
  - HFpEF patients have left ventricular EF > 50%
- HFpEF of great interest
  - it has prevalence of nearly 50% of patients and it is not well understood and there are very few treatments available for the patients
- EMRs/medical claims typically do not contain the HFpEF diagnosis, therefore case ascertainment (phenotyping) of the HF strains is highly desirable

MERCK

Example 1:
Phenotyping Cancer Stage (Proof of Concept)

- Cancer stage is the most important risk factor associated with survival and its ascertainment from the medical claims data is desirable to support RWE
  - Cancers are historically diagnosed at late stage
- Identification of cancer stage from administrative data a major challenge
  - Traditionally used algorithms based on “clinically” defined decision trees deliver poorly on achieving high sensitivity/specificity simultaneously
- Classifier ensembles (superlearner) are showing promise in classification tasks when using with big data such as medical claims

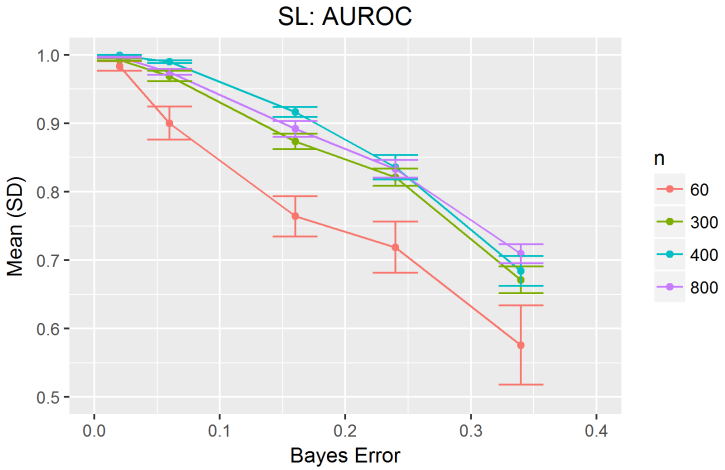
MERCK

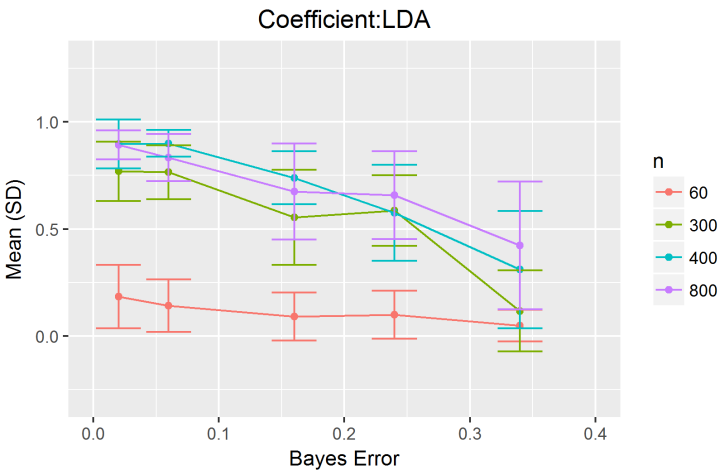
Stacked Generalization (Metalearning)
Superlearner

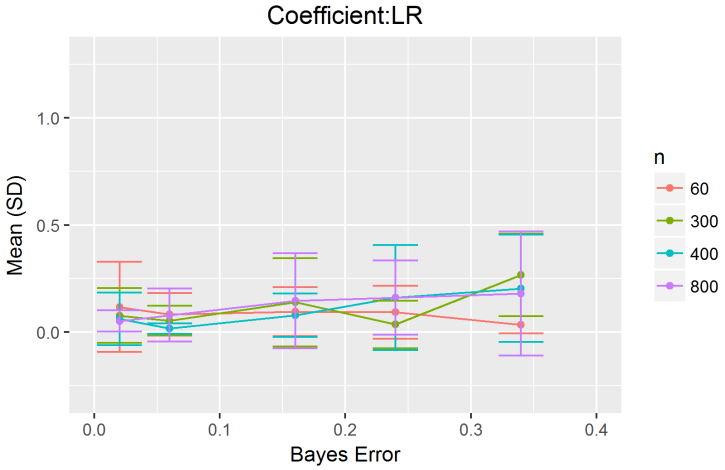
- Level-zero data: The original training set (X, Y)
- Level-one data: The cross-validated predicted values (Z - becoming meta-features)
- Learner Library: Set of basis learners (learning algorithms)
- Meta-learner: Algorithm trained using
  - cross-validated predicted values Z and
  - the original target Y
  - typically linear

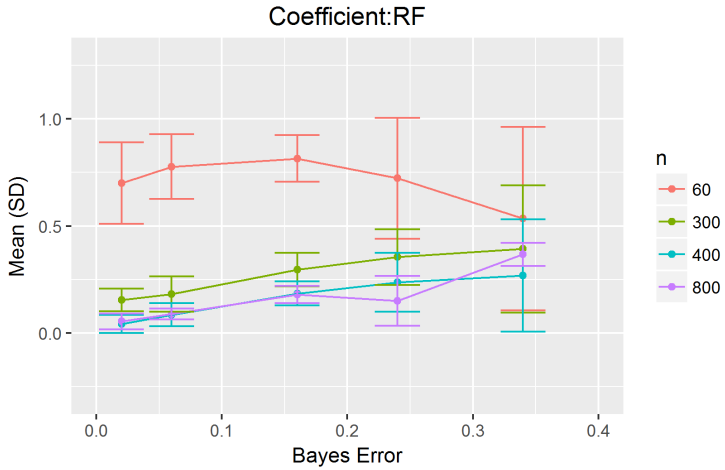
MERCK

Superlearner for Linearly Separable Data







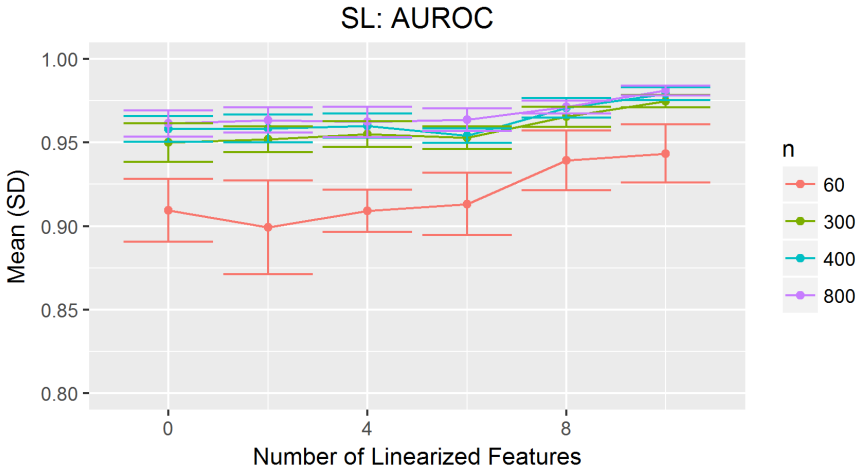


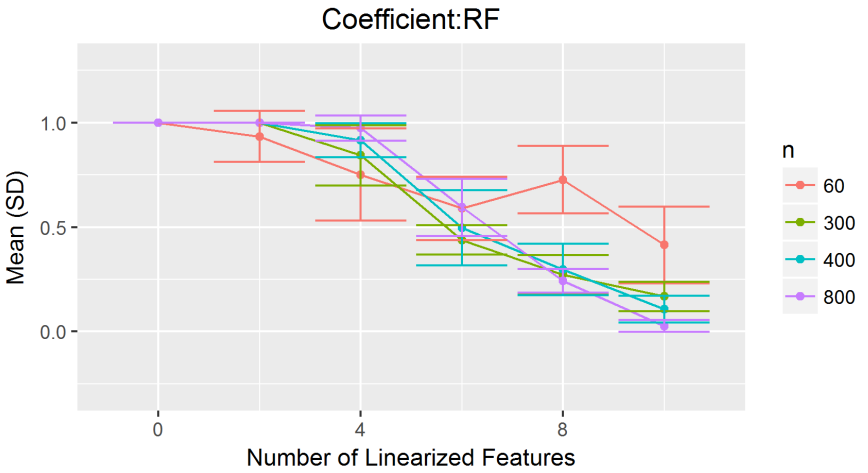
SL Coefficients provide insight into relative merit of component classifiers

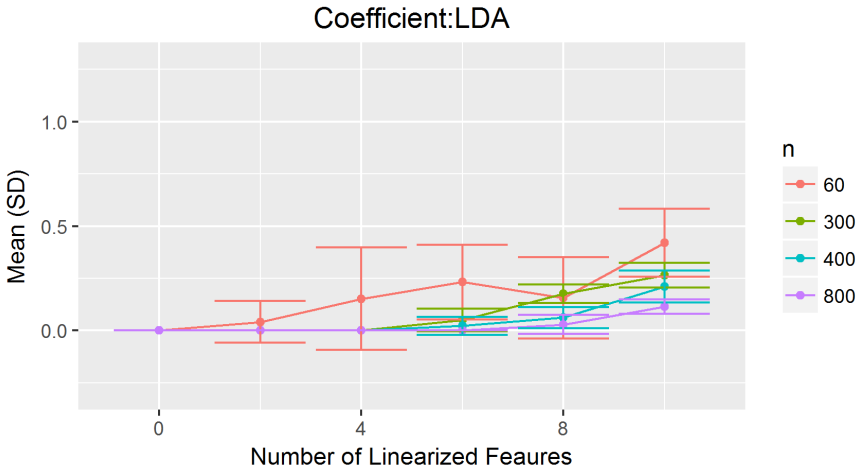
Interesting interplay between Random Forest and LDA

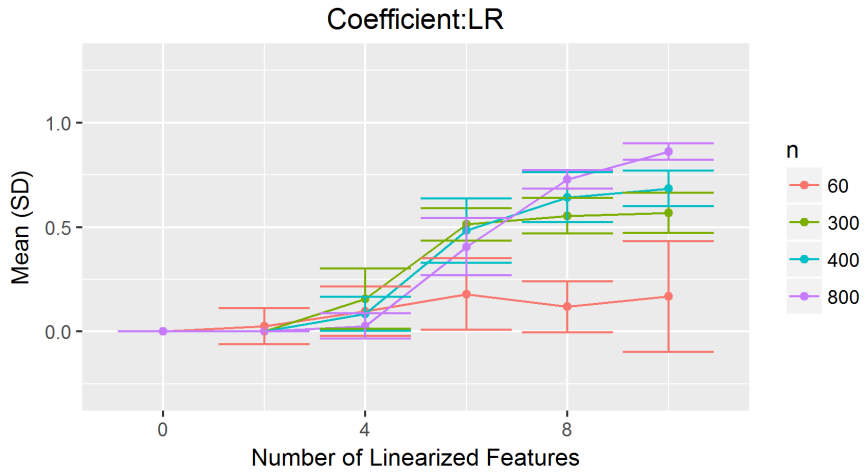
MERCK

Superlearner for non-Linearly Separable Data









SEER and Medicare Linked Data

- A linkage between SEER cancer registry and Medicare administrative claims database
  - SEER is a large cancer registry that covers 25% of the US population
  - Medicare claims cover comprehensive inpatient/outpatient diagnosis and procedures received by Medicare beneficiaries
- Study included patients who were diagnosed with lung cancer and received chemotherapy in 2010-2011
    - Cancer stage classification algorithms were developed from Medicare inpatient/outpatient claims
    - Cancer stage classification from SEER registry served as gold standard for validation
      - Early: AJCC stage I/II/II (local)
      - Late: AJCC stage IV (metastatic)

MERCK

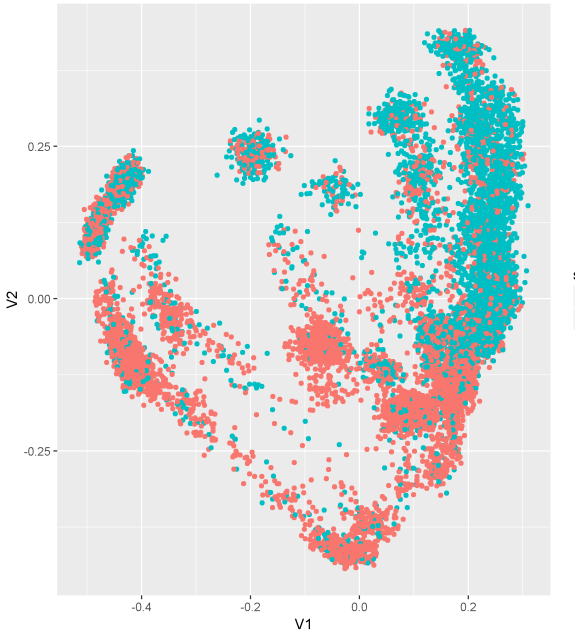
Study Cohort and Data Set Construction

- A total of 11,198 patients were included
- The constructed data set included three sets of predictors (Bergquist et al, 2017)
  - C1: clinical variables (suggested by a clinical tree algorithm)
  - C2: demographics, lung surgery, radiation therapy, chemotherapy regimen, comorbidities
  - C3: lung and secondary malignancies diagnosis
- A total of 101 predictors
  - Qualitative (categorical): 68
  - Quantitative (continuous): 33
- Target class – late stage lung cancer according to SEER (gold standard)
  - Early: AJCC stage I/II/III n=6,039
  - Late: AJCC stage IV n=5,159

MERCK

Results and Visualization by Unspervised Random Forest

All variables projected to two dimensions denoted as V1 and V2



Top 5 variables obtained from random forest

Superlearner comprising of logistic Regression, xgboost and Random Forest achieved balanced sensitivity/specificity ~0.8

- C3_11_198_per: % metastases codes other site
- C24_3_noncranial_cnt: Number of claims for non-brain radiation
- C23_3_lobec: C23_3_lobec
- C3_8_met_cnt: Number of claims for metastases claims
- C3_9_196_per: % metastases codes lymph node site

MERCK

Summary of the Cancer Staging POC

- From the simulations
  - Ensembling by a superlearner shows good performance for larger data sets
  - Coefficients obtained from the superlearner are informative and provide insights about the data geometry
- From the real-world study
  - Current ML approaches significantly outperformed secondary cancer diagnosis
  - Random forest and superlearner exhibited superior performance in terms of sensitivity and specificity with respect to the logistic regression and xgboost
  - Superlearner also provided balanced sensitivity and specificity

MERCK

Ongoing and Future Work

- ECOG (Eastern Cooperative Oncology Group) phenotyping from EMR data (Flatiron data base)
  - Currently ongoing as an analysis data set is being constructed

- Phenotyping of frailty from medical claims data
  - Medicare Current Beneficiary Survey (MCBS) provides linked database of claims and self reported functional status (activity of daily living)

- Phenotyping of heart failure
  - Main challenge: finding appropriate data sources
  - Work on EMR conducted by Rishi Desai's team at the Harvard medical School

MERCK

Broader View on Added Value of Artificial Intelligence (AI) and Statistical Learning in Enrichment of RWD

- Analysis of non-structured heterogeneous data sources with Artificial Intelligence (AI) approaches including natural language processing (NLP) and deep learning is currently intensively researched (Shickel et al. 2018, Liao et al. 2015)
  - Natural Language processing (nlp) for EHR/EMR phenotypes in heart failure patients identification, identification of patients with pancreatic cancer history, etc.
  - Supervised Deep Learning:
    - Long-Short-Term-Memory (LSTM) networks and convolutional neural networks (CNNs) in diagnosis classification from multivariate time-series obtained from EHRs
  - Unsupervised Deep Learning:
    - Autoencoder (Deep Patient) generation of patient representation from EHR data

- Leveraging recent fast emerging developments in relevant AI techniques provides for enrichment of RWD and enables its use towards generation of reliable RWE in early and late stage drug development

MERCK

Implementation Challenges

- **Many Software Available**

  R vs SAS
  - R 15,000 statistical packages
  - R first released the latest cutting edge techniques
  - R less “standard” documentation /too many packages (same function, statistics with many algorithms, packages) /validation
  - R less/no support

  - SAS easy to learn with enormous variety of statistical functions and great GUI (Enterprise Guide & Miner) and with new standard in analytics & data management (like Viya etc)
  - SAS with strong data manipulation functions
  - SAS less ML options, needs additional package beside SATA, like EM for (k-fold-cv) RF, SVM and many High-Performance Procedures etc

- **Chose SAS extracting from big databases, R for ML**

MERCK

Improvement of Running Time (SuperLearner)

- **PC Rstudio**
  - >30 hours
- **R Parallel Package**
  - many different levels, with 'coarse-grained parallelism'
    - library(parallel)
    - library(snow)
    - library(doMC)
- **Requesting CPUs (Cores or Slots)**
  - Default request for the number of CPUs is one
  - "-pe threads 4" (Syntax in Grid Engine)
  - Use parallel and 4 cores in HPC Rstudio (Linux)
    - Time: 1.149801 days (27.6 hours)
    - Use parallel and 8 cores
      - Time: 20.75816 hours
    - Use parallel and 8 cores process (8-way jobs on the same node)
      - Time: 7.99262 hours

MERCK

R Platforms/Tools/Issues

- **PC R, RStudio vs HPC Rstudio (Linux)**
  - Syntaxes/Comments
  - Tools
  - Running time
- **Not Stable of Run Time (R Parallel in Linux)**
  - Known issue using parallel code in R when using fork.
  - need to switch over to use Rmpi which uses the openmpi module

References

- On the use of real-world evidence to support regulatory decision–making for medical devices, FDA draft guidance 2017
- Jarow JP, LaVange L, Woodcock J: Multidimensional Evidence Generation and FDA Decision Making, JAMA, 2017
- Jacqueline Corrigan-Curay, Real world evidence and path forward FDA, Duke Margolis conference, 2017
- Nancy A. Dreyer: Advancing a framework for the use of real-world evidence: when real is reliable, Therapeutic Innovation and Regulatory Science, 2018
- SL Bergquist et al. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. Proceedings of Machine Learning Research, 68:25-38, 2017
- Faurot K et al. Using claims data to predict dependency in activities of daily living as a proxy for frailty. Pharmacoepidemil. Drug Safety, 24(1), 59-66, 2015
- Desai R et al. Algorithms to predict left ventricular ejection fraction class in patients with heart failure using Medicare claims data. ICPE, Prague, 2018.
- Shickel et al.: Deep EHR: A survey of recent advances in deep learning techniques for EHR analysis. IEEE Journal of Biomedical and Health Informatics. 22(5), 1589-1604, 2018.
- Liao et al.: Development of phenotype algorithms using EMRs and incorporating natural language processing. 350, BMJ, 2015.

MERCK

ECOG Performance Status

The ECOG performance status is a scale used to assess how a patient's disease is progressing, assess how the disease affects the daily living abilities of the patient, and determine appropriate treatment and prognosis

- Grade 0: fully active, able to carry on all pre-disease performance without restriction
- Grade 1: restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light housework, office work
- Grade 2: ambulatory and capable of all self-care but unable to carry out any work activities, up and about more than 50% of waking hours
- Grade 3: capable of only limited self-care, confined to bed or chair more than 50% of waking hours
- Grade 4: completely disabled, cannot carry on any self-care, totally confined to bed or chair
- Grade 5: dead

https://radiopaedia.org/articles/ecog-performance-status

MERCK