Challenges to the classic randomized clinical trial

- Up to 75% of trials don’t reach participant enrollment targets*
- A single clinical trial can cost up to $300 million*

- Trial populations often aren’t representative of the general population and typical healthcare situations
- Many trials provide inadequate information on important subgroups because of limited enrollment
- Patients, clinicians, and other users of healthcare data have limited input into trial design and conduct

Big data with linkage potential

- Electronic Medical Records
  - Single system, multi-system
- Insurance claims
  - Medicare, Medicaid, Commercial
- Registries (e.g. SEER)
- Primary data
  - Patient reported outcomes
  - Research-specific assessments
- Lab data, imaging, pathology
- Genomic data
- Aggregate data from geographical units
  - Air pollution, water quality
  - Weather (heat waves)
  - SES
- Wearables
  - Fit bit, smart phone apps
- Social media, internet searches, purchases (pregnancy tests)
Motivating examples

• 81 vs 325mg aspirin and CVD outcomes and GI bleed as major endpoints
• Long term outcomes after bariatric surgery including all-cause mortality
• Risk of cancer between two antidiabetic medications
• Effects of anti-emetics on pregnancy outcomes
• Receipt of recommended care postpartum among moms of medically fragile infants vs. well newborns
• Healthcare utilization among children with autism
Following patients through time and space
Identifying the right population

• Equipoise

• Assessing inclusion & exclusion criteria
  – Indicated for treatment
  – Not already treated (e.g. new users)
  – Not contraindicated (e.g. CKD, pregnancy)
  – No history of the outcome (e.g. prevalent cancers)
Conditions: Single EHR vs. EHR+claims

- Intracranial hemorrhage: 59%
- Ischemic stroke: 44%
- Congestive heart failure: 39%
- Myocardial infarction: 37%
- Hepatotoxicity: 36%
- Pulmonary embolism: 34%
- Major bleeding: 32%
- Acute kidney injury: 31%
- Deep vein thrombosis: 31%
- Depression: 27%
- GI bleeding: 26%

Source: (Lin et al., 2018)
Medications: Single EHR vs EHR+claims

- Antibiotics: 84%
- Hormone therapy: 59%
- Antidepressants: 51%
- Dementia: 50%
- Statins: 49%
- NSAIDs: 46%
- Anticonvulsants: 46%
- Antihypertensives: 45%
- Oral anticoagulants: 45%
- NSAIDs: 42%
- Statins: 42%
- Dementia: 41%
- Antidepressants: 40%
- Hormone therapy: 37%
- Antibiotics: 36%

Source: (Lin et al., 2018)
Treatments

• Medications
  – Prescribed vs filled vs ingested
  – Inpatient vs outpatient vs OTC
  – Use over time (adherence, persistence)

• Devices
  – General (CPT) vs specific (UDI, serial number)
  – Placed, revised, removed
Fraction of EMR prescription medications actually filled

Source: (Fischer et al., 2018)
Outcomes

- Verifying presence of health outcomes
  - Labs, imaging, pathology, clinical notes
- Across care settings
  - Urgent care vs usual provider vs hospital
- Across health systems
  - Mobile populations
- Identifying patients in the risk set (denominator)
  - Restricting to loyal patients vs insurance enrollment dates
- Identifying competing events (e.g. mortality)
Risk factors (potential confounders)

- **Clinical details**
  - BMI, BP, lab results, smoking status, renal function

- **Timing**
  - Before or after exposure (lipid tests and statin exposure)
  - Recent vs distal (MI, cancer)

- **Presence vs severity**
  - Type 2 diabetes dx code vs HbA1c
  - Heart failure dx code vs. ejection fraction
  - Cancer dx vs pathology and stage
To link or not to link
Assess feasibility

• Conditions for use
  – DUA
  – Limitations on linkage
  – Data security requirements, access

• Available identifiers or linkage keys
  – Quality, completeness, uniqueness
Assess the potential gains

- Population overlap
  - Sufficient to be valuable
  - Consistently identifiable to avoid double/triple counting some patients or events
  - Consider selection
Scientific value

• Key data made possible only through linkage

• Data quality
  – Accuracy, completeness, differences in coding practice

• More valid, robust, or precise inference
  – Gold standard or alloy

• One time use vs. future research potential
  – NDI
Governance

- Review process for use of the data
- Requirements for data security
- DUA and limits to linkage
- Patient consent, IRB review
Costs

- Data costs
  - Server / disk space
- Personnel, training
  - Server admin
  - Honest broker
- Complexity, time
Linkage Execution

- Cleaning, standardizing, normalizing
- Linkage approach
  - Probabilistic vs deterministic
- Linkage conduct
  - Investigator, one of the original data holders, or honest broker
- Evaluation and validation of record linkage
  - Gold standard available?
- Reporting results
Linking EHR and Claims Data @ UNC
UNC’s EMR Warehouse: CDW-H

- The Carolina Data Warehouse for Health
- Aggregate of electronic health record data collected in UNCHCS, live as of 2009
- Data on ~5+ M unique patients, 800K+ continuous, expanding with UNCHCS
- Additional hospitals added as they “go live” with Epic
  – If it’s in Epic (or was in WebCIS), it’s in the CDW-H.
- Data collection dates back to:
  – July 2004: Hospital Billed Data
  – July 2008: Physician Billed Data
  – April 2014: Epic Systems Data
CDW-H: Not just Chapel Hill

- UNC Health Care System includes hospitals and clinics across North Carolina
- Epic roll-out started in April, 2014

<table>
<thead>
<tr>
<th>Practice</th>
<th>Epic live date</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNC Hospitals and outpatient</td>
<td>April 4, 2014</td>
</tr>
<tr>
<td>UNC Faculty Physicians</td>
<td>April 4, 2014</td>
</tr>
<tr>
<td>Chatham Hospital</td>
<td>April 4, 2014</td>
</tr>
<tr>
<td>UNC Physicians Network West</td>
<td>April 4, 2014</td>
</tr>
<tr>
<td>Rex Healthcare and outpatient</td>
<td>June 20, 2014</td>
</tr>
<tr>
<td>UNC Physicians Network East</td>
<td>June 20, 2014</td>
</tr>
<tr>
<td>Caldwell Memorial, outpatient only</td>
<td>May 5, 2015</td>
</tr>
<tr>
<td>Johnston Health, outpatient only</td>
<td>May 5, 2015</td>
</tr>
<tr>
<td>High Point Regional, outpatient only</td>
<td>May 5, 2015</td>
</tr>
<tr>
<td>Johnston Health Hospitals</td>
<td>May 21, 2016</td>
</tr>
<tr>
<td>High Point Regional Hospital</td>
<td>May 21, 2016</td>
</tr>
<tr>
<td>Caldwell Memorial Hospital</td>
<td>June 18, 2016</td>
</tr>
<tr>
<td>Pardee Hospital and outpatient</td>
<td>June 18, 2016</td>
</tr>
</tbody>
</table>
i2b2 at UNC

- Self-service, web-based query tool which researchers can pull back deidentified data (counts) from the CDW-H
- Launched in March 2015, used by 600+ users at UNC (and growing) to perform data queries preparatory to research
- i2b2@UNC Requirements:
  - UNC-Chapel Hill faculty, staff, or student with an active ONYEN (ID and password)
  - Attendance at Introduction to i2b2 Training
  - Online training now available
Claims data at UNC

- NC TraCS, Epidemiology, Gillings Innovation Lab, and the Sheps Center support access to a variety of claims data sources:
  - 20% national random sample of FFS Medicare claims (2006-2016)
  - 100% of UNCHCS Medicare FFS patient claims linkable with EMR data
    - ~275K people (2015-2016)
  - NC Medicaid data (100%) ~7 years, updated regularly
  - NC BCBS data (100%) ~7 years, updated regularly
  - SEER-Medicare-Part D
  - NC state discharge data (includes ED, surgery center)
# Insurance claims linkable to UNC EHR data

<table>
<thead>
<tr>
<th></th>
<th>Medicare</th>
<th>NC Medicaid</th>
<th>BCBSNC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>FFS Medicare (~20% &lt;65); seen in UNCHS 2014-2017</td>
<td>Low income; pregnant women</td>
<td>Privately insured, &lt;65</td>
</tr>
<tr>
<td><strong>Total n</strong></td>
<td>280k</td>
<td>2.4m (2018)</td>
<td>475k (2018)</td>
</tr>
<tr>
<td><strong>Approvals required</strong></td>
<td>CMS (reuse)</td>
<td>CCQI / NC DMA</td>
<td>CCQI / BCBCNC</td>
</tr>
<tr>
<td><strong>Identifiers used for linkage</strong></td>
<td>HIC (Medicare ID), (birthdate, gender)</td>
<td>Name, birthday, SSN, zipcode</td>
<td>Name, birthday, SSN, zipcode; procedure(s)</td>
</tr>
<tr>
<td><strong>Linkage execution</strong></td>
<td>CMS (existing crosswalk)</td>
<td>NC DMA</td>
<td>Honest broker (UNC’s Sheps Center)</td>
</tr>
</tbody>
</table>
Stage 1 – Ask the experts

- Consult with TraCS CER/BERD regarding
  - Appropriate data sources
  - Feasibility (including timelines, anticipated sample size)
  - Study design
  - Analytic plan
- Consult with Sheps Center re: linkage process as well as expectations, costs, and timeline for honest broker data linkage work
- Consult w/ TraCS Bioinformatics re: cost of data extraction from CDW-H
Stage 2 - Approvals

- UNC Institutional Review Board

- In parallel, prepare applications for permission to use CDW-H and claims data source(s) from:
  - CDW-H oversight (UNC EHR data)
  - ResDAC / CMS (Medicare)
  - CCQI / NC DMA (NC Medicaid)
  - CCQI / BCBSNC (BCBSNC)

- Submit once IRB approval granted
Stage 3 – Primary cohort identification

- Computable phenotype (algorithm) to be applied to structured data
  - Validated method when possible

- Key data elements needed for linkage, claims extraction
  - Study-specific ID
  - Identifiers (name, insurance type, insurance number, birthdate, zip code, sex)
  - Index date (clinical event, treatment, calendar time)

- EHR-derived data on exposure, outcome, patient characteristics extracted by TraCS analyst
Stage 4 - Linkage

- **NC Medicaid**
  - Identifiers provided to NC DMA
  - Crosswalk between encrypted Medicaid ID and Study ID returned to honest broker at Sheps
  - Claims for linked individuals extracted and provided to research team in project-specific work space on Sheps secure server

- **BCBSNC**
  - Identifiers provided to honest broker at Sheps along with any ‘blocking’ criteria to limit the pool of potential matches
  - Claims for linked individuals extracted and provided to research team in project-specific work space on Sheps secure server

- **Medicare**
  - Patids provided to Medicare programmer
  - Relevant claims extracted and provided in project-specific work space on Sheps secure server.
Stage 5 – Actual research (finally)!

- EHR and claims-derived data placed in project-specific folder within Sheps secure server
- Study’s analytic programmer creates analytic cohort
- Investigator or statistical programmer conducts analysis
- Present, publish, and improve public health
- Remember to cite UNC’s CTSA
Study population options

1. Only those that link (overlap)
   - Akin to ‘complete case’ analysis
   - Smallest n
   - Consider potential biases

2. Claims + partial EMR (gray)
   - Population-based cohort with enhanced ascertainment of clinical details in a subset

3. EMR + partial claims (blue)
   - Health-system based cohort with enhanced ascertainment of outcomes, adherence, co-morbid conditions in a subset
Methods for analyzing partial data

• Sensitivity analysis in subset with full data
  – May differ due to same biases that affect the ‘complete case’ analysis

• Quantitative bias analysis
  – Conduct the main analysis in the primary data
  – Adjust results using estimates of sensitivity / specificity (possibly differential) from the linked sample

• Multiple imputation
  – Requires outcomes in the subsample to perform well

• Propensity score calibration
  – Does not require outcome in the subsample
  – Surrogacy assumption needed
Care4moms

• Comparing access to routine post-partum visit and other recommended healthcare services between mothers of medically fragile infants (NICU >3 days) vs those with well babies
• Deliveries (n=6849) at UNC hospital, 7/2014 - 6/2016
• Linkage to claims data attempted for n=1687
• Context-based blocking (delivery during the relevant time period) and fuzzy matching on combinations of first and last name
• Linkage rate 97%
Bariatric surgery

- Comparison of surgical approaches to treat obesity
- Primary outcomes
  - Change in BMI, improvement in diabetes, reoperation, hospitalization, death
- Linkage to claims to identify subsequent operations and hospitalizations up to 5 years later
- Results using linked claims pending
Take home messages

• Both EHR and claims data have important gaps that can lead to substantial bias in estimated treatment effects
• Combining complementary data from EHR + claims often strengthens studies that would otherwise rely on a single data domain
• Benefits need to be weighed against costs (time, funding, complexity)
• Encourage early discussions to assess both
Questions

Michele Jonsson Funk
mfunk@unc.edu

Request a consultation with CER/BERD at
https://tracs.unc.edu/index.php/consultation
Challenges in claims-centric clinical research

- Unmeasured risk factors (e.g. BMI, smoking status)
- Poorly measured disease severity (e.g. HbA1c, ejection fraction), indications (e.g. depression), contra-indications (renal impairment; allergies; pregnancy)
- Unobservable periods (e.g. medications administered during inpatient stay)
- Inability to conduct chart review to verify cases
Challenges in EHR-centric clinical research

- Missing medications / comorbid conditions from encounters outside of the health system
  - Esp when longitudinal follow-up is needed
- Person-time at risk poorly defined
- Health-system specific practices
  - Protocols that dictate treatment
- Selected patient population
  - Tertiary care hospital vs community hospital
  - Public (accepts Medicaid patients) vs Private (avoids Medicaid insured when possible)
The CDW-H Data Model

- The CDW-H contains data in all of the following domains (and more), BUT no master dictionary:
  - Patient demographics
  - Encounter details
  - Diagnoses
  - Procedures
  - Providers
  - Patient vitals
  - Lab tests
  - Medications
  - Orders
  - Notes
  - Charges and Payors
  - Surgery
  - Labor and delivery
  - Medical and social history
  - Patient-reported data
  - Custom data elements