Big Data in Research
The National Inpatient Sample

Salvador Cruz-Flores, MD, MPH, FCCM
Professor and Founding Chair
Department of Neurology
Disclosures

• Nothing to disclose specific to this presentation
Objectives

• Understand the role of big administrative data in clinical research

• Understand the limitations of big administrative data

• Understand the “how to use” the available big datasets through examples
Outline

- Why big data
- Limitations of big data
- Strengths of big data
- Big data sources
- How to
- Examples
- Our department
Why big administrative data anyway?

• Collecting data for research purposes is:
  • Costly
  • Time consuming
  • Limits sample size, follow up and geographical distribution

• Surveillance requires:
  • Continuous data collection
  • Large geographic areas
  • Over long time
Health Services Administrative Data

Wealth of information

- Demographics
- Diagnoses
- Co morbidities
- Treatment
- Hospital utilization
- Disposition
Health Services Administrative Data

- Routinely generated through interactions with health care system
- Reflect "real-world" practice
- "Secondary" data as not generated for research
- It does requires appropriate "research" questions
Sources of bias

- **Population**
  - Capture of cases
  - Setting of encounter
  - Health care models

- **Documentation**
  - Quality of data in medical record
  - Completeness
Health Services Administrative Data

Sources of bias

• Coding
  o How well ICD 9 ICD 10 reflects clinical encounter
  o How well code reflects the actual condition
  o Financial incentives influence coding

• Data Linkage
  o Incorrect or erroneous matching
  o Delays inherent to compilation of data
<table>
<thead>
<tr>
<th>Documentation</th>
<th>Administrative Codes*</th>
<th>Interpretation of Codes</th>
</tr>
</thead>
</table>
| 85 male with diabetes mellitus, ischemic cardioembolic stroke caused by atrial fibrillation with hemorrhagic transformation | I63.4 (MRDx)  
I48.90 (DxType1)  
E11.52 (DxType3) | Cerebral infarction caused by embolism of cerebral arteries  
Atrial fibrillation, unspecified  
Type 2 diabetes mellitus with certain circulatory complications |
| 85 male, type 2 diabetes mellitus, stroke and hemorrhage on CT, atrial fibrillation | I61.9 (MRDx)  
I48.90 (DxType2)  
E11.52 (DxType3) | Intracerebral hemorrhage, unspecified  
Atrial fibrillation, unspecified  
Type 2 diabetes mellitus with certain circulatory complications |

Scenario: 85-year-old man with diabetes mellitus type 2 admitted with an ischemic stroke with hemorrhagic transformation and is found to have atrial fibrillation on the second day of admission. CT indicates computed tomography; DxType, diagnosis type; and MRDx, most responsible diagnosis.

*These represent real codes generated by a coding specialist in a tertiary-care hospital based on hypothetical patient information.
Health Services Administrative Data

Stroke Surveillance

- Observational epidemiology
- Distribution and determinants of stroke
- Inequities in health and healthcare
- Disease association with risk factors
- Impact of research knowledge translation to large populations overtime
Databases available

CDC datasets

- Behavioral Risk Factor Surveillance System (BRFSS)
- Cancer Data and Statistics
- Chronic Disease and Health Promotion Open Data
- Chronic Disease Indicators (CDI)
- Chronic Disease State Policy Tracking System
- Data, Trends and Maps
- Diabetes Data and Statistics
- Health-Related Quality of Life (HRQOL)
- National Assisted Reproductive Technology Surveillance System (NASS)
- Youth Risk Behavior Surveillance System (YRBSS)
Databases available

- Agency for Healthcare Research and Quality (AHRQ)
  - Healthcare Cost and Utilization Project (HCUP)
    - [https://www.hcup-us.ahrq.gov/databases.jsp](https://www.hcup-us.ahrq.gov/databases.jsp)
    - National (Nationwide) Inpatient Sample (NIS)
    - Kids’ Inpatient Database (KID)
    - Nationwide Emergency Department Sample (NEDS)
    - Nationwide Readmissions Database (NRD)

- Premier Database
Research question

• Are there racial-ethnic disparities in hospital utilization among patients with intracerebral hemorrhage?

• Exposure of interest:
  o Racial/Ethnic group vs Whites

• Outcomes of interest:
  o Therapeutic procedures: Ventriculostomy, Craniotomy, VP shunt
  o Life sustaining procedures: mechanical ventilation, tracheostomy, gastrostomy, transfusion
  o Palliative care, comfort measures,
  o In hospital mortality
Research question

• Covariates and adjustments
  o Co-morbidities
  o Insurance status
  o Severity scores
Racial-Ethnic Disparities in Hospital utilization in patients with Intracerebral hemorrhage

Table 1: Demographic, clinical characteristics, in-hospital events and outcomes of patients with intracerebral hemorrhage (ICH) across different ethnicities in United States (NIS: 2008-2014).

<table>
<thead>
<tr>
<th></th>
<th>White*</th>
<th>Black*</th>
<th>Hispanic*</th>
<th>Other*</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>227752</td>
<td>59848</td>
<td>32669</td>
<td>30171</td>
<td></td>
</tr>
<tr>
<td>Age (mean years ± SD)</td>
<td>71.1 (±14.9)</td>
<td>61.0(±14.8)</td>
<td>62.5 (±17.5)</td>
<td>65.5 (±16.2)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Women</td>
<td>115419 (50.7)</td>
<td>29364 (49.0)</td>
<td>14079 (43.1)</td>
<td>13972 (46.3)</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Racial-Ethnic Disparities in Hospital utilization in patients with Intracerebral hemorrhage

Multivariate analysis determining the effect of race/ethnicity strata on in-hospital therapeutic procedures among patients with intracerebral hemorrhage

<table>
<thead>
<tr>
<th>Races</th>
<th>Ventriculostomy™ OR (95%CI), p value</th>
<th>Cranietomy™ OR (95%CI), p value</th>
<th>Ventriculoperitoneal shunting™ OR (95%CI), p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>White race</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>Black race</td>
<td>1.154 (0.968- 1.376), p = 0.1111</td>
<td>1.090 (0.694- 1.713), p = 0.7071</td>
<td>1.269 (0.921- 1.749), p = 0.1458</td>
</tr>
<tr>
<td>Hispanic race</td>
<td>1.092 (0.978- 1.220), p = 0.1164</td>
<td>1.282 (0.981- 1.674), p = 0.0684</td>
<td>1.157 (0.898- 1.491), p = 0.2600</td>
</tr>
<tr>
<td>Other race</td>
<td>1.027 (0.955- 1.104), p = 0.4750</td>
<td>1.049 (0.853- 1.291), p = 0.6480</td>
<td>1.096 (0.926- 1.296), p = 0.2874</td>
</tr>
</tbody>
</table>
Racial-Ethnic Disparities in Hospital utilization in patients with Intracerebral hemorrhage

Multivariate analysis determining the effect of race/ethnicity strata on in-hospital procedures among patients with intracerebral hemorrhage

<table>
<thead>
<tr>
<th>Races</th>
<th>Mechanical ventilation</th>
<th>Gastrostomy</th>
<th>Transfusion</th>
<th>Tracheostomy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95%CI), p value</td>
<td>OR (95%CI), p value</td>
<td>OR (95%CI), p value</td>
<td>OR (95%CI), p value</td>
</tr>
<tr>
<td>White</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>Black</td>
<td>1.145 (1.068- 1.227),</td>
<td>1.427 (1.307- 1.559),</td>
<td>0.946 (0.866- 1.034),</td>
<td>1.358 (0.924- 1.996),</td>
</tr>
<tr>
<td></td>
<td>p = 0.0001</td>
<td>p = &lt;.0001</td>
<td>p = 0.2205</td>
<td>p = 0.1190</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1.045 (1.004 - 1.088),</td>
<td>1.117 (1.058- 1.180),</td>
<td>1.043 (0.990- 1.098),</td>
<td>1.208 (0.970- 1.505),</td>
</tr>
<tr>
<td></td>
<td>p= 0.0326</td>
<td>p= &lt;.0001</td>
<td>p= 0.1108</td>
<td>p= 0.0910</td>
</tr>
<tr>
<td>Other</td>
<td>1.071 (1.041- 1.101),</td>
<td>1.133 (1.090- 1.177),</td>
<td>1.008 (0.976- 1.042),</td>
<td>1.289 (1.132- 1.466),</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.0001</td>
<td>p = &lt;.0001</td>
<td>p = 0.6188</td>
<td>p = 0.0001</td>
</tr>
</tbody>
</table>
Racial-Ethnic Disparities in Hospital utilization in patients with Intracerebral hemorrhage

Multivariate analysis determining the effect of race/ethnicity strata on in-hospital therapeutic procedures among patients with intracerebral hemorrhage

<table>
<thead>
<tr>
<th>Races</th>
<th>In-hospital mortality*</th>
<th>Palliative care*</th>
<th>Do not resuscitate (DNR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95%CI), p value</td>
<td>OR (95%CI), p value</td>
<td>OR (95%CI), p value</td>
</tr>
<tr>
<td>White</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>Black</td>
<td>0.798 (0.746- 0.854),</td>
<td>0.664 (0.601- 0.732),</td>
<td>0.658 (0.588- 0.735),</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.0001</td>
<td>p= &lt;.0001</td>
<td>p= &lt;.0001</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.875 (0.842- 0.910),</td>
<td>0.817 (0.765- 0.872),</td>
<td>0.951 (0.890- 1.015),</td>
</tr>
<tr>
<td></td>
<td>p= &lt;.0001</td>
<td>p= &lt;.0001</td>
<td>p= 0.1306</td>
</tr>
<tr>
<td>Other</td>
<td>0.929 (0.905- 0.953),</td>
<td>0.907 (0.870- 0.947),</td>
<td>0.941 (0.897- 0.987),</td>
</tr>
<tr>
<td></td>
<td>p= &lt;.0001</td>
<td>p= &lt;.0001</td>
<td>p= 0.0123</td>
</tr>
</tbody>
</table>
Conclusions

• Racial minorities have similar utilization of therapeutic “curative” procedures as compared to Whites

• Racial minorities have greater utilization of life sustaining therapies as compared to Whites

• Racial minorities have a lower in hospital mortality and lower utilization of palliative care service and comfort measures as compared to Whites
Is Thrombolysis Safe in the Elderly?
Analysis of a National Database

Amer Alshekhlee, MD, MSc; Afshin Mohammadi, MD; Sonal Mehta, MD; Randall C. Edgell, MD; Nirav Vora, MD; Eli Feen, MD; Sushant Kale, MD; Zaid A. Shakir, MD; Salvador Cruz-Flores, MD, MPH

Background and Purpose—Thrombolysis for acute ischemic stroke in the elderly population is seldom administered.

Methods—In this study, we evaluated the risks of thrombolysis, including the mortality and intracerebral hemorrhage (ICH) rates in this population. A cohort of patients was identified from the National Inpatient Sample database for the years 2000–2006. Age was categorized in 2 groups, including those between 18 and 80 years and those >80 years. Multivariate logistic regression analysis was used to assess covariates associated with hospital mortality and ICH. A total of 524,997 patients were admitted for acute ischemic stroke; 143,093 (27.2%) were >80 years. A total of 7950 patients were treated with thrombosis, of which 1659 (20.9%) were >80 years. Elderly patients received less frequent thrombolysis compared with the younger population (1.05% versus 1.72%).

Results—In the whole cohort, the mortality rate was higher in the older population (12.80% versus 8.99%). For those treated with thrombolysis, the mortality rate and risk of ICH were higher among those >80 years (16.9% versus 11.5%; odds ratio: 1.56 [95% CI: 1.35 to 1.82] and 5.73% versus 4.40%; odds ratio: 1.31 [95% CI: 1.03 to 1.67], respectively). Multivariate logistic regression analysis showed that the presence of ICH (odds ratio: 2.24 [95% CI: 1.89 to 2.65]) was associated with higher mortality rates but not the use of thrombolysis (odds ratio: 1.14 [95% CI: 0.98 to 1.33]).

Conclusions—Despite the higher mortality rate in the older population, the use of thrombolysis does not predict death; however, the use of thrombolysis was associated with high risk of ICH. (Stroke. 2010;41:2259-2264.)
Table 2. Bivariated Analysis for Outcomes in Elderly Patients Treated With Thrombolysis Compared With Those <80 Years (Treated With Thrombolysis) and Compared With the Older Population (Untreated With Thrombolysis)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Age &gt;80 y (n=1659)</th>
<th>Age 18 to 80 y (n=6291)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All patients treated with thrombolysis, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital mortality</td>
<td>280 (16.9)</td>
<td>721 (11.5)</td>
<td>1.56 (1.35 to 1.82)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>95 (5.73)</td>
<td>277 (4.40)</td>
<td>1.31 (1.03 to 1.67)</td>
</tr>
<tr>
<td>Endotracheal intubation</td>
<td>139 (8.38)</td>
<td>790 (12.56)</td>
<td>0.63 (0.52 to 0.76)</td>
</tr>
<tr>
<td>Gastrointestinal hemorrhage</td>
<td>22 (1.33)</td>
<td>72 (1.14)</td>
<td>0.89 (0.52 to 1.52)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>17 (1.02)</td>
<td>72 (1.14)</td>
<td>1.16 (0.71 to 1.87)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thrombolysis Yes (n=1659)</th>
<th>Thrombolysis No (n=156921)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All patients age &gt;80 y, n (%)</td>
<td></td>
</tr>
<tr>
<td>Hospital mortality</td>
<td>280 (16.9)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>95 (5.73)</td>
</tr>
<tr>
<td>Endotracheal intubation</td>
<td>139 (8.38)</td>
</tr>
<tr>
<td>Gastrointestinal hemorrhage</td>
<td>22 (1.33)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>17 (1.02)</td>
</tr>
</tbody>
</table>

Graph: Percent (%) of outcomes over years from 2000 to 2006 by hospital type (All elderly patients, Teaching hospitals, Non-teaching hospitals, Urban hospitals, Rural hospitals).
Hospital Mortality and Complications of Electively Clipped or Coiled Unruptured Intracranial Aneurysm

Amer Alshekhlee, MD, MSc; Sonal Mehta, MD; Randall C. Edgell, MD; Nirav Vora, MD; Eli Feen, MD; Afshin Mohammadi, MD; Sushant P. Kale, MD; Salvador Cruz-Flores, MD, MPH

Background and Purpose—To determine the hospital mortality rates associated with elective surgical clipping and endovascular coiling of unruptured intracranial aneurysms.

Methods—We identified a cohort of patients electively admitted to US hospitals with the diagnosis of unruptured intracranial aneurysm from the National Inpatient Sample database for the years 2000 through 2006. Patient demographics, hospital-associated complications, and in-hospital mortality were compared among the treatment groups. A multivariate logistic regression analysis was used to identify independent variables associated with hospital mortality. Cochrane–Armitage test was used to assess the trend of hospital use of these procedures.

Results—After data cleansing, 3738 (34.3%) patients had aneurysm clipping and 3498 (32.1%) had endovascular coiling. The basic demographics including age, race, and comorbidity indices were similar between the groups. The length of hospital stay was longer in the clipped population (median 4 versus 1 day; P<0.0001), incurring a higher hospital charge in the coiled population (median $42 070 versus $38 166; P<0.0001). Hospital mortality was higher in the clipped population: 60 (1.6%) versus 20 (0.57%; adjusted odds ratio 3.63; 95% CI, 1.57, 8.42). Perioperative intracerebral hemorrhage and acute ischemic stroke were higher in the clipped population. The rate of hospital use of the endovascular coiling has increased over the years included in this study (<0.0001).

Conclusions—Elective coiling of unruptured intracranial aneurysms is associated with fewer deaths and perioperative complications compared with elective clipping. The trend of hospital use of the coiling procedures has increased during recent years. (Stroke. 2010;41:1471-1476.)
### Table 3. Multivariate Logistic Regression Analysis Shows Independent Variables Associated With Treatment Type

<table>
<thead>
<tr>
<th>Effect</th>
<th>OR (95% CI)</th>
<th>P Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Died (yes vs no)</td>
<td>3.63 (1.57, 8.42)</td>
<td>0.002</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50 vs &gt;75 y</td>
<td>4.28 (3.01, 6.07)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>51 to 75 vs &gt;75 y</td>
<td>4.79 (3.39, 6.78)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Female vs male</td>
<td>1.43 (1.26, 1.63)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Postoperative stroke (yes vs no)</td>
<td>1.73 (1.27, 2.35)</td>
<td>0.0005</td>
</tr>
<tr>
<td>Sepsis (yes vs no)</td>
<td>4.06 (0.88, 18.72)</td>
<td>0.07</td>
</tr>
<tr>
<td>Pulmonary (yes vs no)</td>
<td>1.51 (0.86, 2.64)</td>
<td>0.14</td>
</tr>
</tbody>
</table>

*Hosmer–Lemeshow goodness-of-fit test for the model shows $P=0.74$ with maximum inflation of 1.19.

---

**Graph**: Clipped vs Coiled

- **Clipped**
  - y2000: 569
  - y2001: 475
  - y2002: 458
  - y2003: 498
  - y2004: 602
  - y2005: 376
  - y2006: 560

- **Coiled**
  - y2000: 0
  - y2001: 49
  - y2002: 711
  - y2003: 711
  - y2004: 777
  - y2005: 645
  - y2006: 1040

**UIA Clipped**
- Total = 3738

**UIA Coiled**
- Total = 3448
### Table 6. Diagnostic and Procedure Codes Used to Identify and Treat Ruptured and UIAs in National Databases

<table>
<thead>
<tr>
<th></th>
<th>Barker et al(^3)</th>
<th>Qureshi et al(^13)</th>
<th>Cowan et al(^14)</th>
<th>Current Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 clinical modification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unruptured aneurysm</td>
<td>437.3</td>
<td>437.3</td>
<td>437.3</td>
<td>437.3</td>
</tr>
<tr>
<td>Ruptured aneurysm</td>
<td>430.0</td>
<td>430.0</td>
<td>430.0/431.0*</td>
<td>None</td>
</tr>
<tr>
<td>Procedure codes for aneurysm repair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clipping</td>
<td>39.51</td>
<td>39.51</td>
<td>39.51</td>
<td>None</td>
</tr>
<tr>
<td>Wrapping/endovascular</td>
<td>39.52†</td>
<td>39.52†</td>
<td>39.52†</td>
<td>None</td>
</tr>
<tr>
<td>Pure endovascular</td>
<td>None</td>
<td>None</td>
<td>39.72/39.79‡</td>
<td>39.72/39.79‡</td>
</tr>
</tbody>
</table>

*Authors included intracerebral hemorrhage along with subarachnoid hemorrhage.
†Procedure code 39.52 is given to other type of aneurysm repair including surgical treatment such as coagulation, suturing, wiring, and wrapping, used for aneurysm coil repair.
‡Procedure codes 39.72 and 39.79 include coil embolization or occlusion of aneurysm as well as endovascular grafting. Cowan et al used these codes for patients admitted in 2001 through 2003; and 39.52 was used for aneurysm coil repair before 2001.
Neurology Projects

- Racial disparities in hospital utilization in subarachnoid hemorrhage
- Racial disparities in hospital utilization in vegetative state
- Impact of seizures and status epilepticus in outcome ischemic stroke, ICH and SAH
- Outcome Guillain Barre treated with IVIG, Plasma exchange or both
- Readmissions for ICH among patients with ischemic stroke and atrial fibrillation
### Table: Comparison of Big Science With Other Clinical Care Research Approaches

<table>
<thead>
<tr>
<th></th>
<th>Little Science</th>
<th>Big Data</th>
<th>Big Science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>To meet needs of Food and Drug Administration regulations and to generate new knowledge</td>
<td>To support the needs of decision makers</td>
<td>To support the needs of decision makers</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Estimates efficacy on primary end point</td>
<td>Estimates associations, explore hypotheses, draw causality inferences about comparative effects of alternative treatments, effects in subgroups, effects on intermediate or delayed outcomes</td>
<td>Estimate comparative effectiveness measured on outcomes that matter to patients, including burden of alternative courses of action</td>
</tr>
<tr>
<td><strong>Questions</strong></td>
<td>Address a goal important to the sponsor, regulator, or the investigator</td>
<td>Address what is possible given the available data</td>
<td>Address what is possible with generous collaboration and necessary for confident clinical decision making</td>
</tr>
<tr>
<td><strong>Participants</strong></td>
<td>Selected to optimize responsiveness and support favorable indications</td>
<td>Real-world patients receiving all or part of their care in participating health systems or care funded by participating payers and for whom data and follow-up is sufficiently complete</td>
<td>Patients receiving care in practice research networks</td>
</tr>
<tr>
<td><strong>Comparisons</strong></td>
<td>Often placebo</td>
<td>Usual practice</td>
<td>Usual care with evidence-based alternatives that patients value</td>
</tr>
<tr>
<td><strong>Data and sources</strong></td>
<td>Per protocol</td>
<td>As-is in existing data sets</td>
<td>Per protocol</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>Smallest fastest trials; multicenter with few participants per center; and meta-analyses of these trials</td>
<td>Use data collected from linked observational databases formed by the practice and administration of healthcare</td>
<td>Largest possible study (multicenter trials and prospective meta-analyses) to precisely estimate effects, beneficial, and harmful, across a broad range of patients and contexts</td>
</tr>
<tr>
<td><strong>Tactics to achieve efficiency</strong></td>
<td>Enroll responsive participants: use surrogate and composite endpoints, compare to placebo</td>
<td>Ask multiple questions against data collected in the course of providing or paying for care</td>
<td>Use data generated and collected in the course of providing (documented in records) or receiving care (clinician and patient collected/ reported)</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td>Disease-oriented, surrogate markers assessed as soon as possible to obtain significant differences between arms</td>
<td>Surrogate markers, procedures, patient outcomes at sufficiently long follow-up periods by which data is sufficiently complete</td>
<td>Outcomes that decision makers (often patients) experience and value at sensible follow-up periods</td>
</tr>
<tr>
<td><strong>Size/funding relationship</strong></td>
<td>As big as funding allows</td>
<td>As big as available in datasets</td>
<td>As big as possible, requiring multiple funding sources</td>
</tr>
<tr>
<td><strong>Stakeholder involvement</strong></td>
<td>Optional, directed at making funding and recruiting feasible</td>
<td>Optional, directed at selecting analyses and disseminating results</td>
<td>Essential, from identifying question to disseminating results</td>
</tr>
<tr>
<td><strong>Scope of collaboration</strong></td>
<td>Across researchers with outreach to nonresearch clinicians and patient groups to optimize recruitment</td>
<td>Across researchers and data owners/stewards</td>
<td>Across funders, healthcare practice networks, patient groups, researchers and dissemination agents, often internationally</td>
</tr>
<tr>
<td><strong>Confidence in the body of evidence</strong></td>
<td>Indirect evidence in terms of patients, comparators, and outcomes; limited precision, particularly outside of primary end point (pooling can improve but cannot overcome other biases); biased publication related to interest of sponsor and other contacts</td>
<td>Direct and highly precise comparisons are likely, but hindered by residual confounding. Blurred lines between protocol-driven work and data-driven protocols. Publication and reporting bias highly likely</td>
<td>Direct and highly precise results, even across a range of patients and contexts, are likely. Publication bias is less likely thanks to public protocols and visible large-scale conduct in real world</td>
</tr>
</tbody>
</table>
**Funding**
- Internal
  - Institutions & Hospitals
  - Professional societies
- External
  - Venture capital
  - Incubators and Accelerators
- * Crowdfunding

**Training in Data Sciences**
- New educational curriculums
- Structured mentoring in clinical informatics
- Practice modules on standardized datasets
- Open-Access online courses

**Research Teams**
- Clinical investigators
- Bioinformatics
- Engineers
- Healthcare administrators
- Healthcare policy
- Business
- Entrepreneurs

**Reward**
- New metric system based on innovation
- Dedicated new journals for data sciences
- Peer Review
- White Papers
- Social Media
- Intellectual Property
- Novel University career and promotion pathways
Take home

• Big databases can be a powerful instrument for research
• They are relatively cheap
• They can help create pilot data to support grant applications
• They can also create new research questions
• Their use for scientific inference requires an understanding of their strengths and limitations
• First step is to ask the right question to the right database