What we heard

We are missing things
- Race
- "Uncoded drugs"
- Coding validity
- Coverage

We have many terminologies
- ICD-9, ICD-10
- Precision: "Not otherwise specified"
- Procedure drugs

We have complex outcomes
- Liver injury with no prior history
- Database linking
- PPV determinations

We have varying data provenance
- Standard: Diagnoses, pharmacy claims
- Smoking, alcohol use
- Income quintiles
- Death
- Inpatient vs outpatient vs ER

Data

We lack data control /access
- Distributed databases
- No knowledge of data characteristics prior to study

Evidence Based Medicine

- ## AHRQ evidence levels
  - Level I: randomized controlled trial (RCT)
  - Level II-1: controlled trials without randomization
  - Level II-2: uncontrolled cohort or case-control studies
  - Level II-3: Multiple time series with or without the intervention
  - Level III: Opinions of respected authorities

Requirements for Data

- How are data characterized?
  - Delta to complete patient:
    - Intervention (drugs)
    - Outcomes (diagnoses or complex definition)
    - Confounders
- What is data quality?
  - Counts, expected counts, trends over time
  - Validation: of cases - or of tools and scripts
  - Source record verification, PPV
  - Terminologies – precision, coverage, compatibility, ability to represent desired cohorts
- How are cohorts defined?
  - Exposure, Health Outcome of Interests, SMQ, Inclusion/exclusion criteria
  - Definition: transparent, reproducible, independent from study