Claims Data

Medical Claims

Diagnosis, Procedures,

Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor,

Medication Cost

- Electronically available
- Standardized

- Not 100% accurate
- Under-reporting is common
- Claims for hospital visits can be vague

Creating the Dataset – Claims Samples

Claims Sample

- Large health insurance claims database
- Randomly selected 131 diabetes patients
- Ages range from 35 to 55
- Costs $10,000 – $20,000
- September 1, 2003 – August 31, 2005

Creating the Dataset – Expert Review

Claims Sample

Expert Review

- Expert physician reviewed claims and wrote descriptive notes:

“Ongoing use of narcotics”

“Only on Avandia, not a good first choice drug”

“Had regular visits, mammogram, and immunizations”

“Was given home testing supplies”

Creating the Dataset – Expert Assessment

Claims Sample

Expert Review

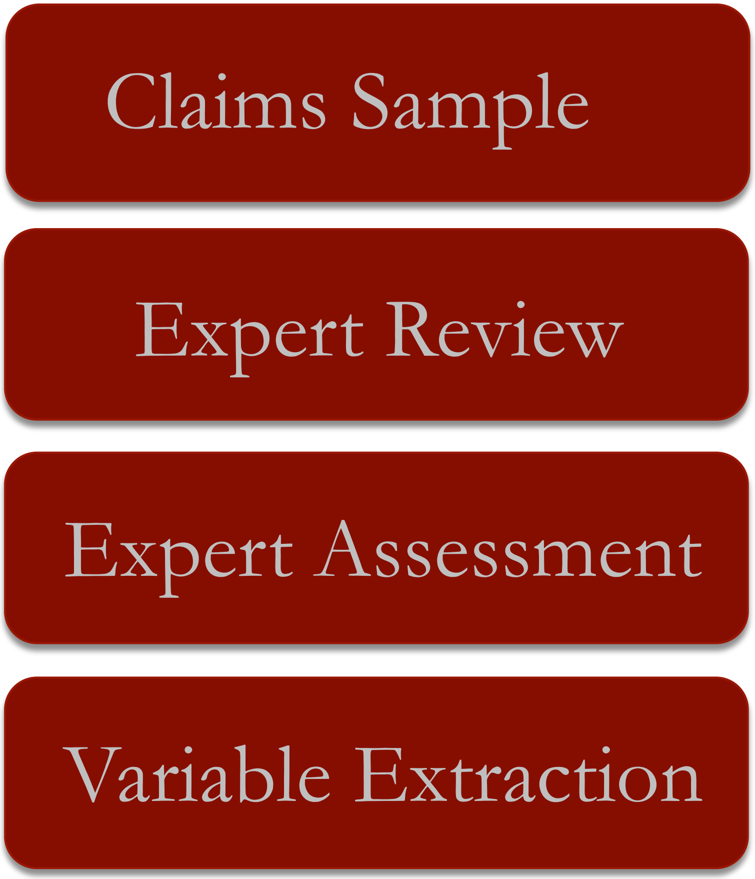
Expert Assessment

- Rated quality on a two-point scale (poor/good)

"I'd say **care was poor** – poorly treated diabetes"

"No eye care, but overall I'd say **high quality**"

Creating the Dataset – Variable Extraction



Claims Sample

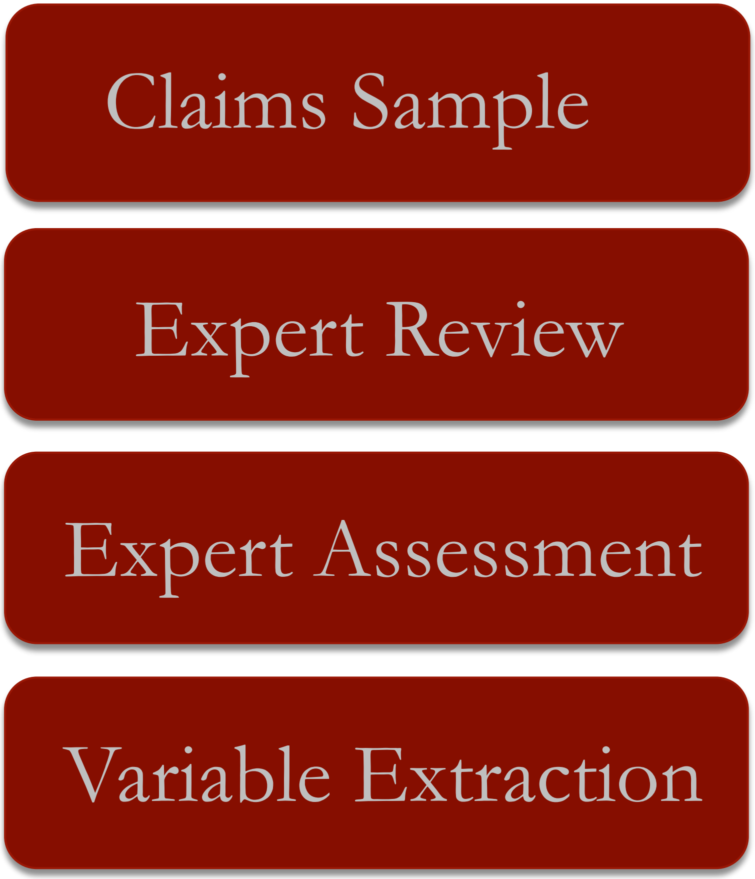
Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
  - **Quality of care**
- Independent Variables
  - ongoing use of **narcotics**
  - **only on Avandia**, not a good first choice drug
  - Had **regular visits, mammogram, and immunizations**
  - Was given **home testing supplies**

Creating the Dataset – Variable Extraction



Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
  - **Quality of care**
- Independent Variables
  - Diabetes treatment
  - Patient demographics
  - Healthcare utilization
  - Providers
  - Claims
  - Prescriptions

Predicting Quality of Care

- The dependent variable is modeled as a binary variable
  - 1 if low-quality care, 0 if high-quality care
- This is a *categorical variable*
  - A small number of possible outcomes
- Linear regression would predict a continuous outcome
- How can we extend the idea of linear regression to situations where the outcome variable is categorical?
  - Only want to predict 1 or 0
  - Could round outcome to 0 or 1
  - But we can do better with logistic regression