Workshop Introduction
Putting the Pieces Together: Precision Medicine Discovery from Electronic Health Records

Janina M. Jeff PhD, MS
Icahn School of Medicine at Mount Sinai
September 22, 2015

EMR PHENOTYPING: WORKING GROUPS

- Body Mass Index
- Age at Menopause
- Age at Natural Menopause
- Type 2 Diabetes
- Myocardial Infarction
- Smoking Status

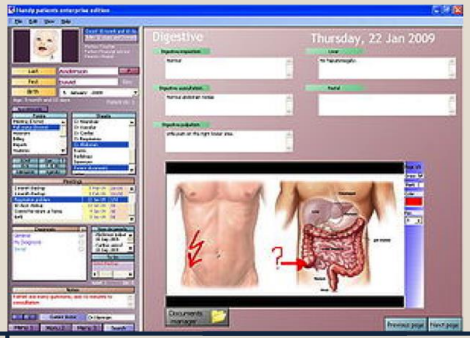
BODY MASS INDEX (BMI)

$$BMI = \frac{weight(kg)}{height\ 2\ (\text{meters})} \quad \text{or} \quad BMI = \frac{\text{weight(pounds)*703}}{\text{height2(in)}}$$

What are some things to consider?

BMI

- Time period restrictions
  - BMI is routinely recorded at each in- and out-patient visit



Median BMI from 12 month period

3 6 9 12

Months

BMI : INCLUSION/EXCLUSION

- Inclusions
  - Adult (18 yrs older)
  - Multiple BMI recordings > 12 month period

- Exclusions based on ICD9 and CPT codes
  - Bariatric surgery
  - Pregnancy
  - Thyroid abnormalities
  - Weight loss/gain medications
  - Diabetes medications/status
  - Extreme Obesity
  - Eating Disorders

BMI : INCONSISTENCIES

- Temporal Inconsistencies
  - Exclude extreme measurements > 3 SD from BMI recordings over 1 year
  - Use other measurements to correct for errant measures

- Unit Inconsistencies
  - pounds to kilograms
  - inches to centimeters
  - feet to centimeters
  - meters to centimeters

AGE AT MENOPAUSE

- **Age at termination of menstruation**
- **Extracting Age (X) (numerals only)**
  - Age at diagnosis of menopause
  - If X is a date, calculate the age in years by subtracting the subject's birthdate from X.
  - If more than one age X is identified, use the age listed most frequently.
  - If more than one age X is identified and recorded an equal number of times, use the first instance of X found in the subject's record.

AGE AT MENOPAUSE

Inclusion:

- Has any of the surgical menopause ICD-9 codes **OR**
- Has any of the surgical menopause CPT codes **OR**
- Has any of the menopause **OR** surgical keywords/pattern matching by free text data mining

- **Primary Exclusion:**
  - Male gender
  - Age ≤ 18 years
  - Has a diagnosis code for Fragile X syndrome (ICD-9 759.83)

- **Secondary Exclusion:**
  - X < 18 or X > 65

AGE AT MENOPAUSE

Inclusion:

- Has any of the surgical menopause ICD-9 codes **OR**
- Has any of the surgical menopause CPT codes **OR**
- Has any of the menopause **OR** surgical keywords/pattern matching by free text data mining

- **Primary Exclusion:**
  - Male gender
  - Age ≤ 18 years
  - Has a diagnosis code for Fragile X syndrome (ICD-9 759.83)

- **Secondary Exclusion:**
  - X < 18 or X > 65

AGE AT NATURAL MENOPAUSE

- **Age at termination of menstruation without surgical or medicinal intervention**
- **Extracting Age (X) (numerals only)**
  - Age at diagnosis of menopause
  - If X is a date, calculate the age in years by subtracting the subject's birthdate from X.
  - If more than one age X is identified, use the age listed most frequently.
  - If more than one age X is identified and recorded an equal number of times, use the first instance of X found in the subject's record.

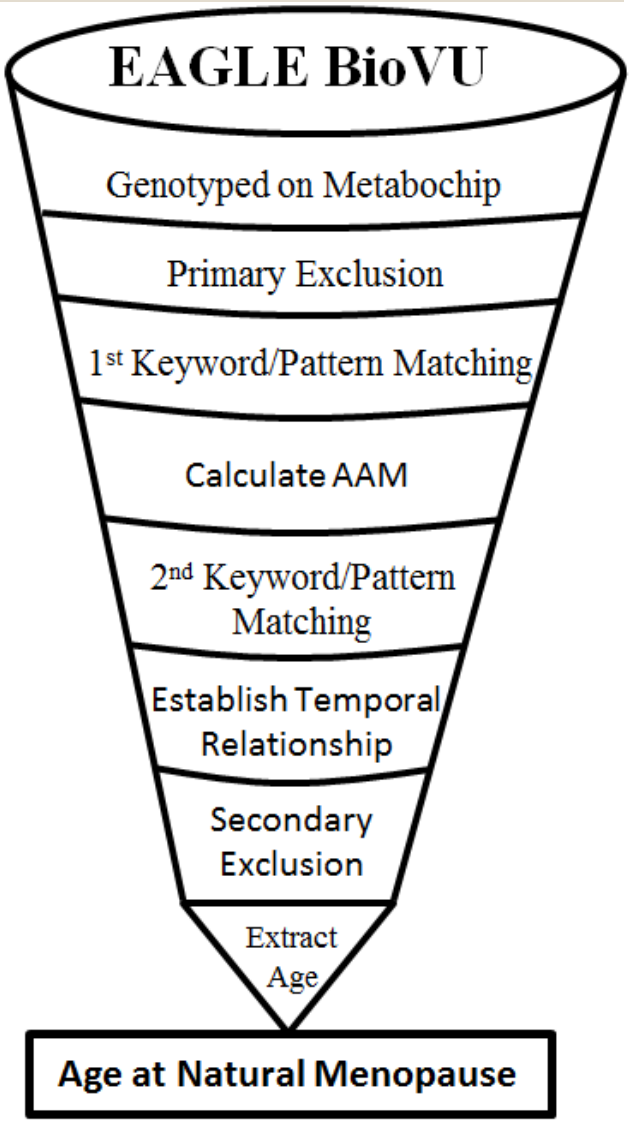
AGE AT NATURAL MENOPAUSE

- **Inclusion:**
  - **TEXT MINING ONLY**
  - Has any of the menopause keywords/pattern matching by free text data mining
- **Menopause keywords/pattern matching:**
  - menopause at X
  - menopause at age X
  - menopause aged X
  - age X at menopause
  - menopause at the age of X
  - menopause – X
  - menopause was at X
  - X at menopause
  - age at menopause: X
  - menopause began at X
  - menopause began at age X

Where X indicates numbers and dates as **numerals** only.

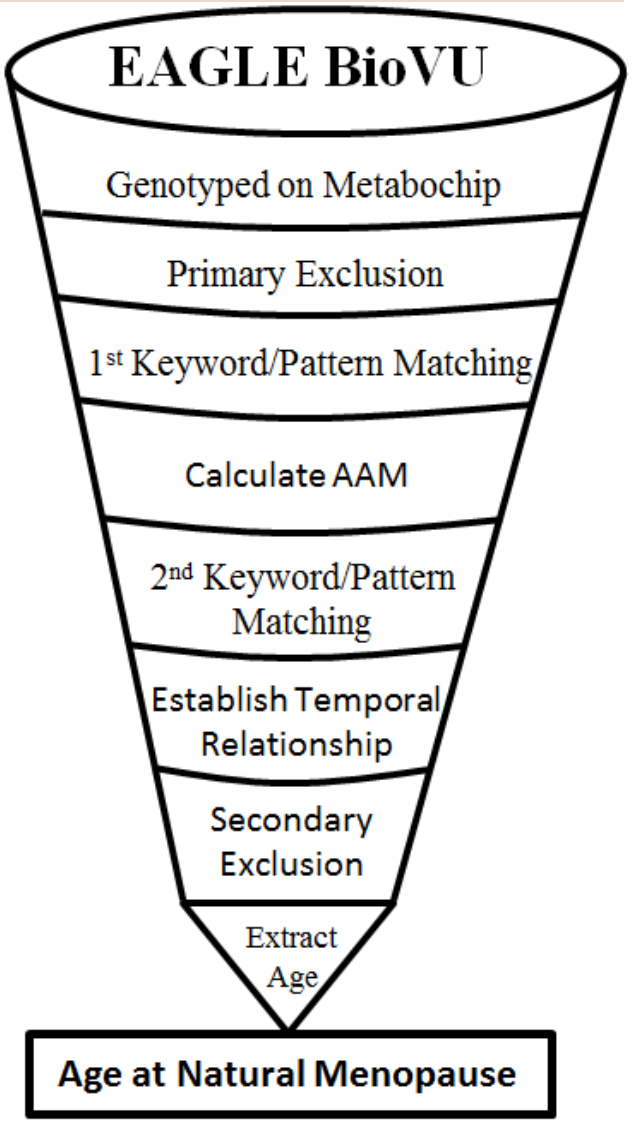
AGE AT NATURAL MENOPAUSE

- **Primary Exclusion:**
  - Male gender
  - Age < 41 years
  - Has a diagnosis code for premature ovarian failure/premature menopause (ICD-9 256.31) **OR** ovarian failure (ICD-9 256.39) **OR** artificially induced menopause (ICD-9 627.4) **OR** Fragile X syndrome (ICD-9 759.83)



AGE AT NATURAL MENOPAUSE

- **Secondary Exclusion:**
  - Has any of the surgical menopause ICD-9 codes **OR** surgical menopause CPT codes **OR surgical** keywords by free text data mining occurring **PRIOR TO AGE X**
  - Has had any of the HRT medications prescribed **PRIOR TO AGE X**
- **Surgical keywords:**
  - surgical menopause
  - total abdominal hysterectomy
  - TAH
  - complete hysterectomy
  - TAH-BSO
  - Oophorectomy
  - laparoscopic hysterectomy
  - uterine ablation
  - endometrial ablation
  - Thermoablation
  - vaginal hysterectomy
  - vaginal radical hysterectomy
- **Tertiary Exclusion**
  - X < 18 or X > 65



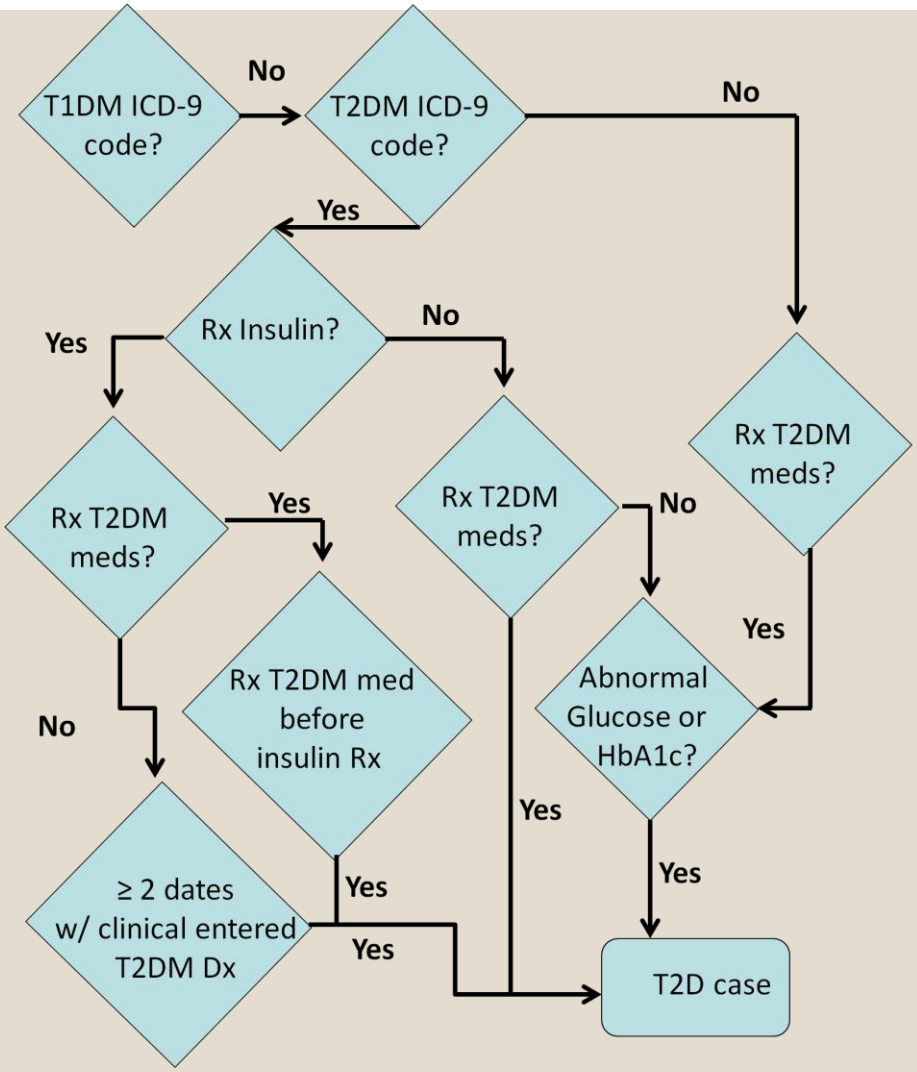
AGE AT MENOPAUSE ACCURACY

- Manual chart review
  - 2x2 tables, randomized participants, 100 with an age at event (50 exact, 50 de-identified), 100 without age at event by algorithm
  - calculation of sensitivity, specificity, and positive predictive value (PPV)

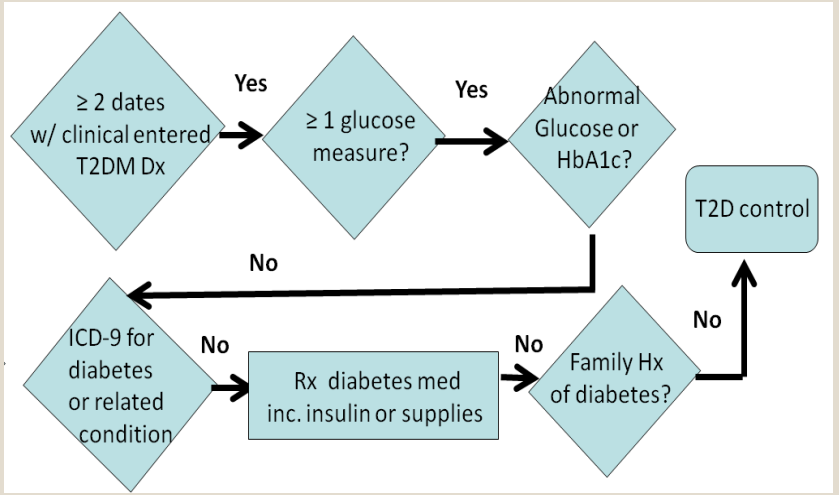
	+ Menopause	- Menopause	Total
+ Algorithm menopause	84	16	100
- Algorithm menopause	5	95	100
Total	89	111	200

  - Accuracy calculated (required EXACT concordance between manual review and algorithm)

Type 2 Diabetes



Case definition



Control definition

MYOCARDIAL INFARCTION

- Complex phenotype
- In vs. Out patient (temporal phenotype?)
- Multiple Case Definitions
- Step-down Approach

MYOCARDIAL INFARCTION

Case 1	At least one MI ICD9-billing codes on 3 consecutive days
Case 2	At least one MI ICD9-billing codes on 2 consecutive days
Case 3	More than 3 MI ICD9-billing codes ever
Case 4	More than 2 MI ICD9-billing codes ever
Case 5	More than 1 MI ICD9-billing codes ever

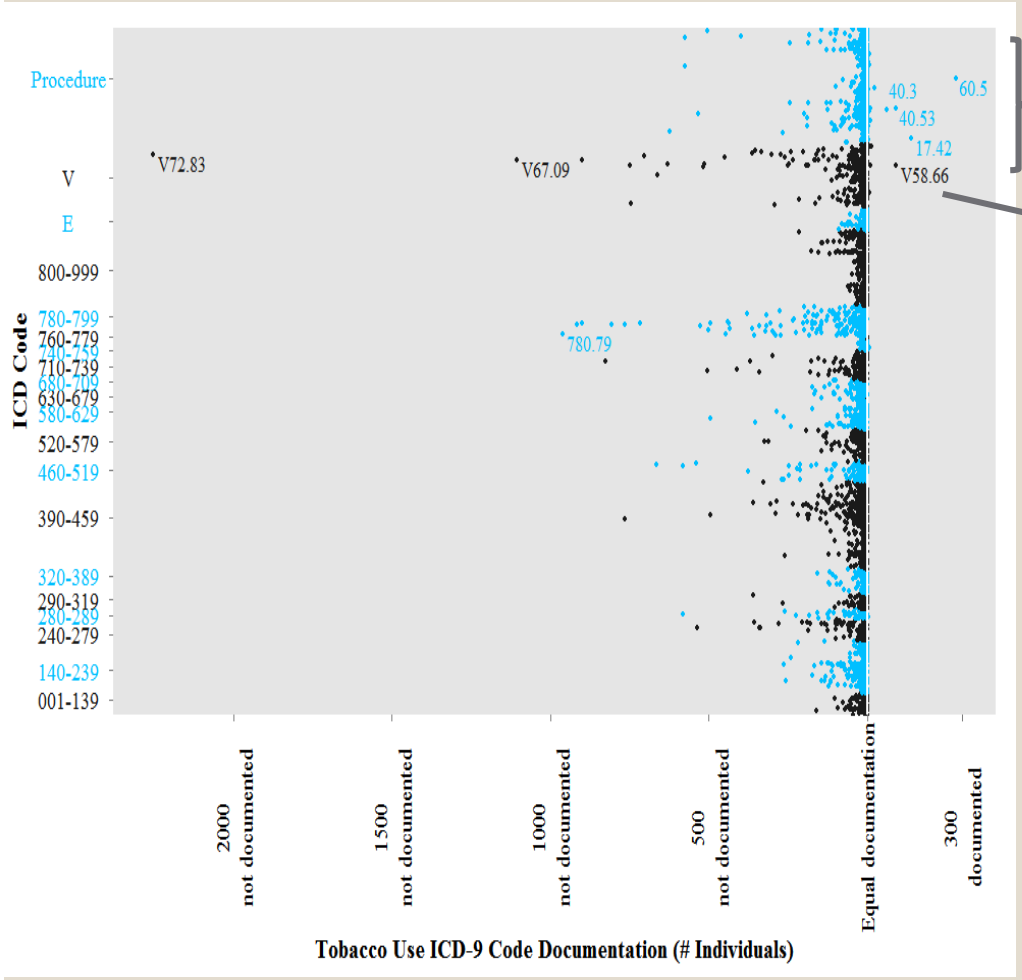
MYOCARDIAL INFARCTION: ACCURACY

Case 1	At least one MI ICD9-billing codes on 3 consecutive days	99.1 %
Case 2	At least one MI ICD9-billing codes on 2 consecutive days	99.4 %
Case 3	More than 3 MI ICD9-billing codes ever	99.4 %
Case 4	More than 2 MI ICD9-billing codes ever	98.1 %
Case 5	More than 1 MI ICD9-billing codes ever	87.6 %

SMOKING STATUS

- Co-Occurrence phenotyping
- Two primary strategies have been applied:
  - Natural Language Processing to scan clinical free text
  - Examination of structured elements/billing code
- ICD9 codes:
  - 305.1 – Tobacco Use Disorder
  - V15.82 – History of Tobacco Use

CO-OCCURRENCE PHENOTYPING SMOKING STATUS



Surgical procedures

Long-term Aspirin Use

Smoking alters the antiplatelet effect of aspirin?
Billing habits differ across clinics?

SMOKING STATUS:ACCURACY

- Through manual review, we established a gold standard set of 100 ever-smokers and 100 never-smokers

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
ICD only	0.32 (0.23-0.41)	1	0.66 (0.59-0.73)
NLP only	0.78 (0.70-0.86)	0.88 (0.82-0.94)	0.83 (0.78-0.88)
ICD + NLP[1]	0.82 (0.75-0.90)	1	0.91 (0.87-0.95)

[1]Ever-smokers if either ICD or NLP (or both) classify as ever smoker.

- ICD9 codes are highly specific, but not sensitive
- The combination of ICD9 codes and NLP works best

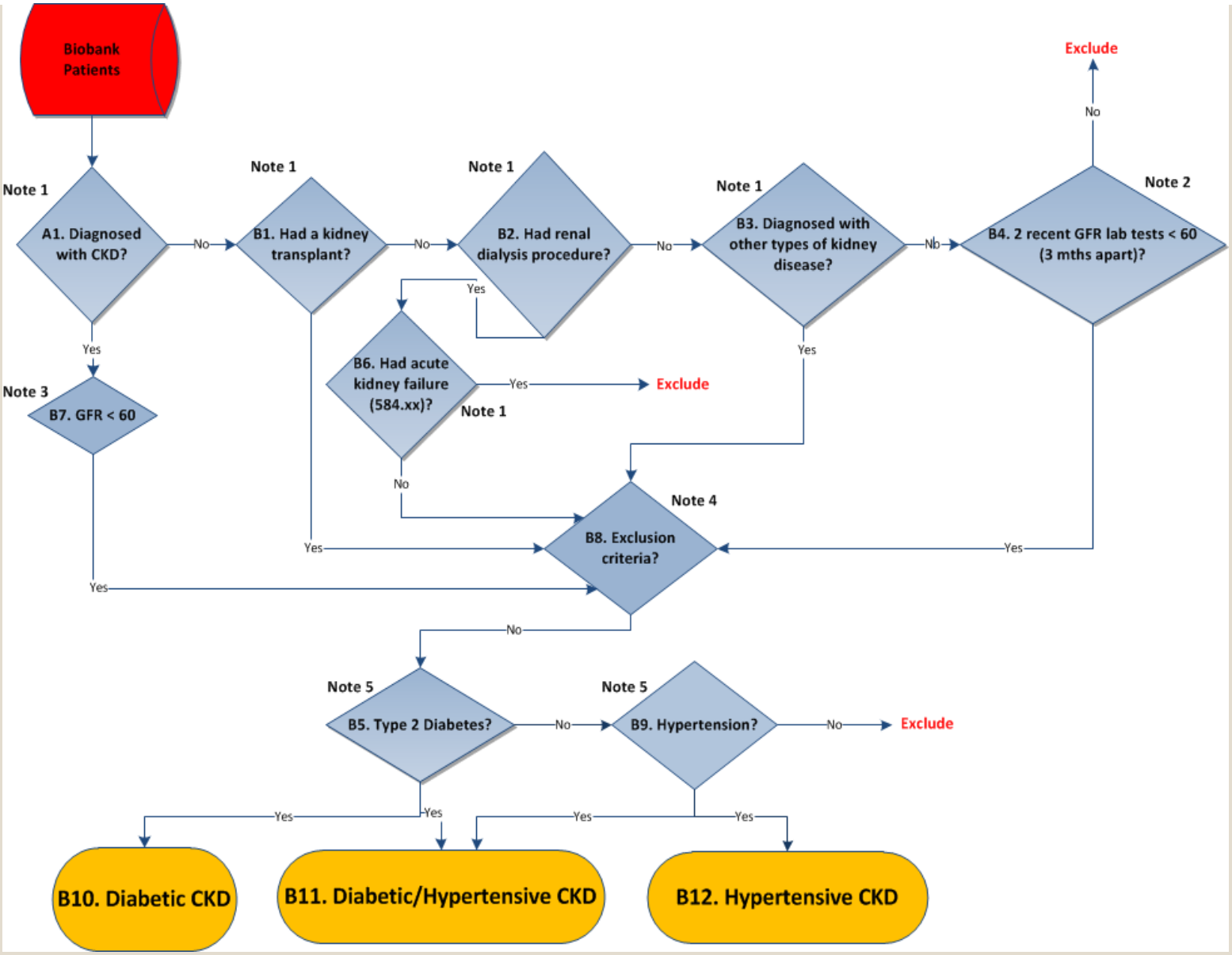
EHR Research Application

CHRONIC KIDNEY DISEASE

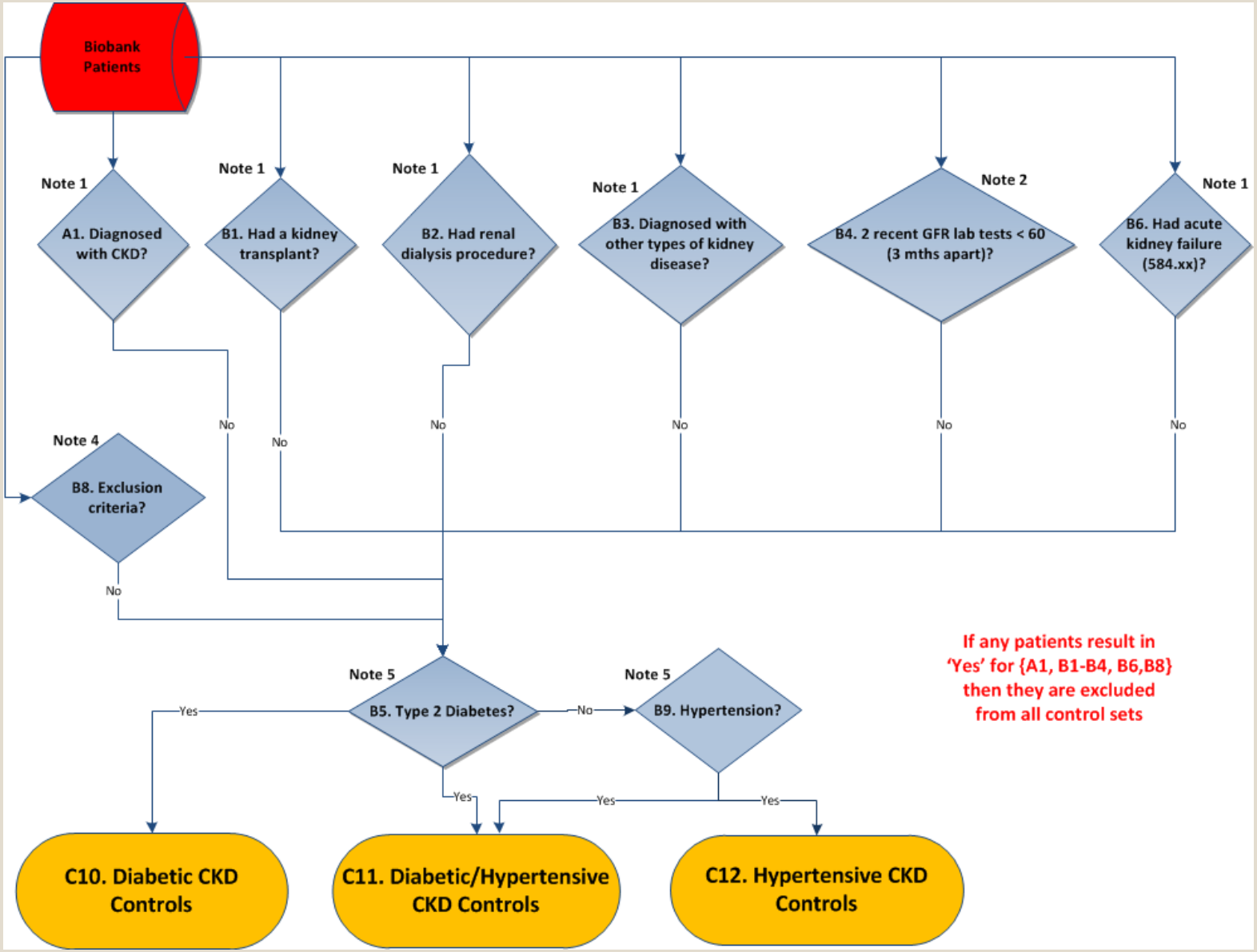
- Complex phenotypes
- Multiple Case/Control Approach
  - Diabetic Kidney Disease
  - Diabetic and Hypertensive Kidney Disease
  - Hypertensive Kidney Disease

- Chronic Kidney Disease is very complex with several possible case and control definitions that are specific to subclinical phenotypes

CHRONIC KIDNEY DISEASE



CHRONIC KIDNEY DISEASE



CODE TRANSITIONS MAY REFLECT CESSATION ATTEMPTS

- ICD9 Codes for Smoking have a temporal component "i.e. History"
- Temporal transitions in the record from "Tobacco Use Disorder" to "History of Tobacco Use" may reflect cessation attempts

	Continuous Smoking	Single Successful Quit Attempt	>1 Unsuccessful Quit Attempts
Single Code Transition	2	25	21
Multiple Code Transitions	8	9	45
$\chi^2 = 18.3725$; df=2; $p<0.0001$			