Duke Medicine

MURDOCK STUDY

The MURDOCK Study: Building a Community-based Registry & Biorepository for Disease Reclassification

Jessie Tenenbaum, PhD
Associate Director for Bioinformatics
Duke Translational Medicine Institute
August 19, 2014

What is the MURDOCK Study?
Measurement to Understand the Reclassification of Disease Of Cabarrus / Kannapolis

- A longitudinal health study to identify genomic linkages across major chronic diseases and disorders
- Organized as *Horizons*
  - Centered in Kannapolis/Cabarrus County, NC
- PI, MURDOCK Study: **Robert M. Califf, MD**
- Co-PIs, MURDOCK Study Community Registry and Biorepository**: L. Kristin Newby, MD, MHS**; **Rowena Dolor, MD, MHS**

- Participating Healthcare and Community Advisory Boards

MURDOCK STUDY

Design: MURDOCK Study Horizons

Horizon 1

- Generated molecular biosignatures using legacy samples, mined associated clinical data
- Generated hypotheses to reclassify diseases: Liver disease, CV disease, Osteoarthritis, Obesity

Horizon 2

- Prospective cohort studies
  - Multiple sclerosis
  - Alzheimer's disease
  - Physical functioning
  - Severe acne
  - Centenarians
- Drawing from subjects accessible via Horizon 1.5 registry

Horizon 3+

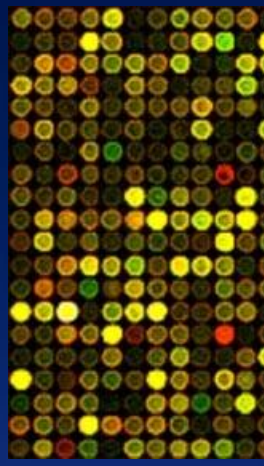
- New waves of molecular data from diverse patient populations
- New cohort studies
- Data meta-analyses
- Multi-institutional and multi-national collaborations
- Measurement of public health impact

Horizon 1.5

- **Building a Community Registry and Biorepository (Informatics and Sample Management)**

MURDOCK STUDY

Horizon One Cohorts



- Cardiovascular Disease
- Hepatitis C
- Obesity
- Osteoarthritis

Horizon 1.5: A Population Registry & Biorepository

Kannapolis, North Carolina

Locally based on the North Carolina Research Campus

Recruitment started in 2009

Open enrollment for 18 and older

Today

>10,600 participants and follow-up
~370,000 samples banked
Median follow-up: ~3 years

Participation in the MURDOCK Registry

- **Inclusion criteria***
  - Age 18 or older**
  - Resides in qualifying zip code OR
  - Male healthcare worker in catchment area regardless of residence

*Study-specific waiver available

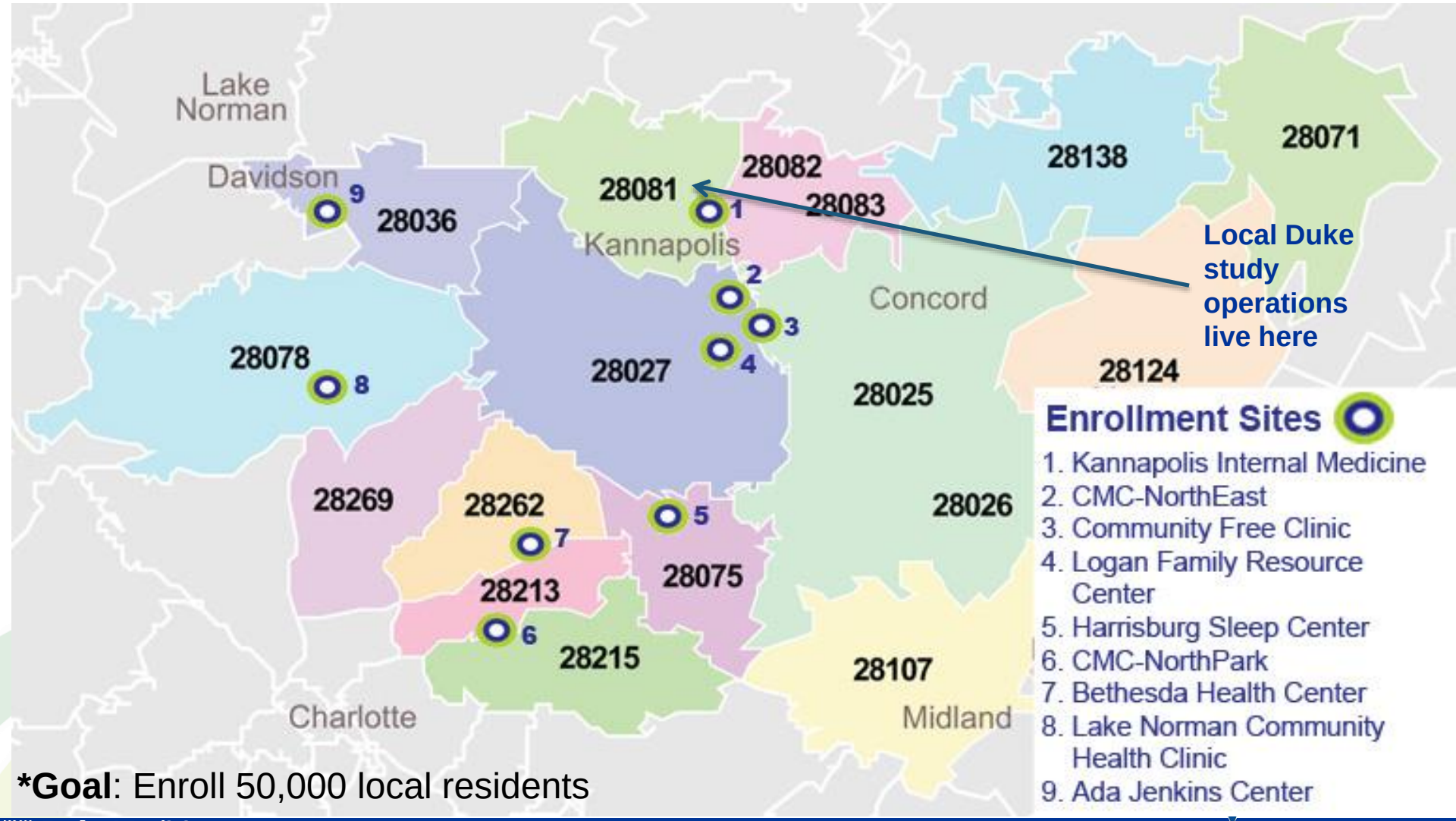
**12 and older in specific cohort recruitment

- **Participants consent to:**
  - ✓ Health and demographic questionnaire
  - ✓ Ongoing access to medical record
  - ✓ Brief exam (HR, BP, waist circumference)
  - ✓ Blood and urine samples for approved research investigations
  - ✓ ***Contact for annual follow-up and up to 4 times yearly to request participation in additional research studies***
  - ✓ Participation indefinite or until consent withdrawn

MURDOCK STUDY

Eligible Zip Codes of MURDOCK Catchment Area

includes Kannapolis and Cabarrus County, and portions of Mecklenburg, Rowan and Stanly Counties, North Carolina



***Goal**: Enroll 50,000 local residents

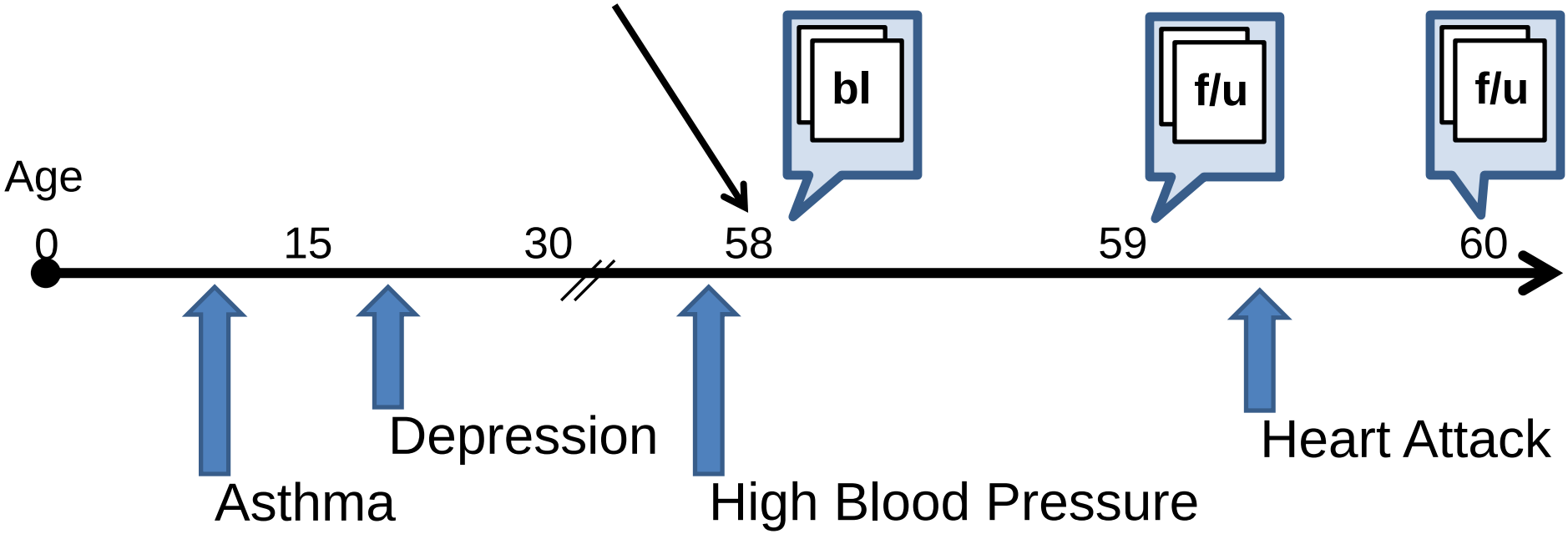
Duke Clinical Research Team

- Locally based, Duke-employed clinical research team
  - 20+ full-time personnel
  - Certified clinical research coordinators, clinical trials assistants, project managers, regulatory staff, and data technicians
  - Spanish-speaking coordinators
  - Translation services
  - On-the-ground recruiters closely engaged with local community

MURDOCK STUDY

Our Health Over Time: A Record Collection

MURDOCK Study enrollment



bl = baseline f/u = follow-up

Duke Translational Medicine Institute

Self-Reported Baseline Data Collected

Participant Baseline Questionnaire completed by each study participant during his/her enrollment visit collects: Demographics, socioeconomic/education/housing, environmental exposures, history of diseases, medications, activity and lifestyle habits, diet and patient-reported outcomes using PROMIS® instrument

*Self-report history of disease data collected on 34 conditions:

Heart	**Metabolic**	**Neurological**
Coronary artery disease	Diabetes	Alzheimer's disease
Atrial fibrillation	High cholesterol	Depression
Heart attack or angina	Thyroid disease	Other mental illness
Congestive heart failure	High blood pressure	Stroke
ICD or pacemaker placement	Obesity	Multiple sclerosis
Cancer	**Lung/Respiratory**	**Gastrointestinal/Renal**
Breast cancer	Asthma	Crohn's disease/ulcerative colitis
Colon cancer	Emphysema or "COPD"	Liver disease
Lung cancer	**Bone/Joint**	Kidney disease
Prostate cancer	Osteoarthritis	
Cervical cancer	Rheumatoid arthritis	
Melanoma	Other autoimmune disease	
Skin cancer, not melanoma	Osteoporosis/Osteopenia	
Oral cancer	Gout	
Other type of cancer		

*Each participant is also geospatially mapped

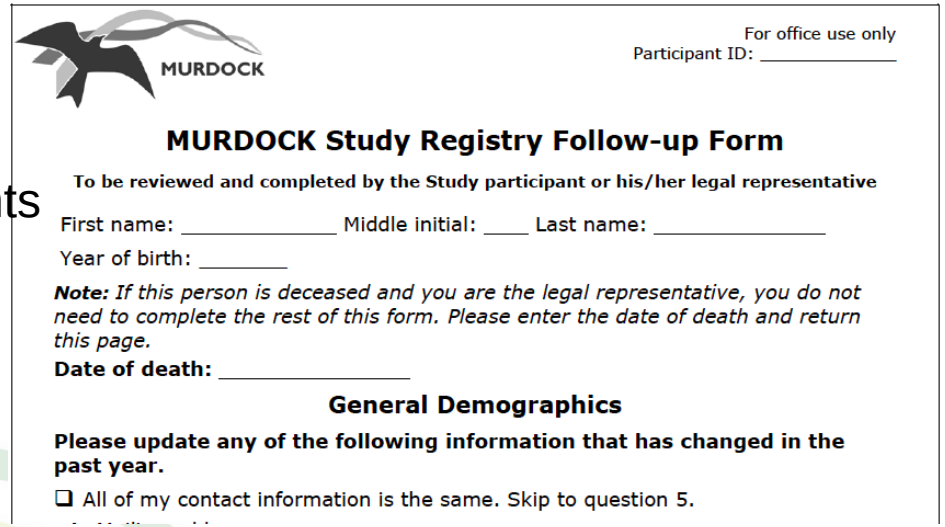
N=10,627 as of 08 Jul 2014

MURDOCK STUDY

Follow-Up Data Collected

Annual Follow Up Forms collect:

- Demographics, changes to medical conditions (same 34 conditions)
  - including procedures, hospitalizations, over previous 12 months during anniversary month of baseline enrollment
- Patient-reported measures using PROMIS® instrument



MURDOCK

For office use only
Participant ID: _____

MURDOCK Study Registry Follow-up Form

To be reviewed and completed by the Study participant or his/her legal representative

First name: _____ Middle initial: ____ Last name: _____

Year of birth: _____

Note: *If this person is deceased and you are the legal representative, you do not need to complete the rest of this form. Please enter the date of death and return this page.*

Date of death: _____

General Demographics

Please update any of the following information that has changed in the past year.

❑ All of my contact information is the same. Skip to question 5.

Activities to promote follow-up:

- Detailed workflow to contact participants
- New study opportunities
- Opportunity to complete follow-up by phone, hard copy, or online (under development)
- Quarterly raffle for completed follow-up
- Quarterly newsletters, birthday and holiday cards, coupons
- Special events

MURDOCK STUDY

Sample Inventory

Average number of specimens obtained per participant: ~35 aliquots
Total N ~370,000 aliquots banked at LabCorp®-Kannapolis Biorepository

Locally stored samples, CAP-accredited, long-term storage facility for Duke efforts

Sample Type	Aliquots/volunteer
Plasma	16 x 500 µL
Buffy coat	from 2 EDTA tubes
Serum	10 x 500 µL
Whole blood	3 x 2 mL
Environmental (Serum)	1 x 3 mL
PaxGene RNA	3 tubes
Urine	4 x 10 mL

LabCorp
Laboratory Corporation of America

- **Engaged participants eager to participate in research**
  - Contact up to 4 times per year to participate in new studies
- **Flexible opportunities and support for collaboration**

MURDOCK STUDY

Snapshot of MURDOCK Study Participants

	% MURDOCK[1]	% Region[2]
Female	65.6	51.3
African American	13.3	25.1
Asian	0.6	3.5
Hispanic	12.0	10.9

Most commonly self-reported diseases / conditions

High cholesterol

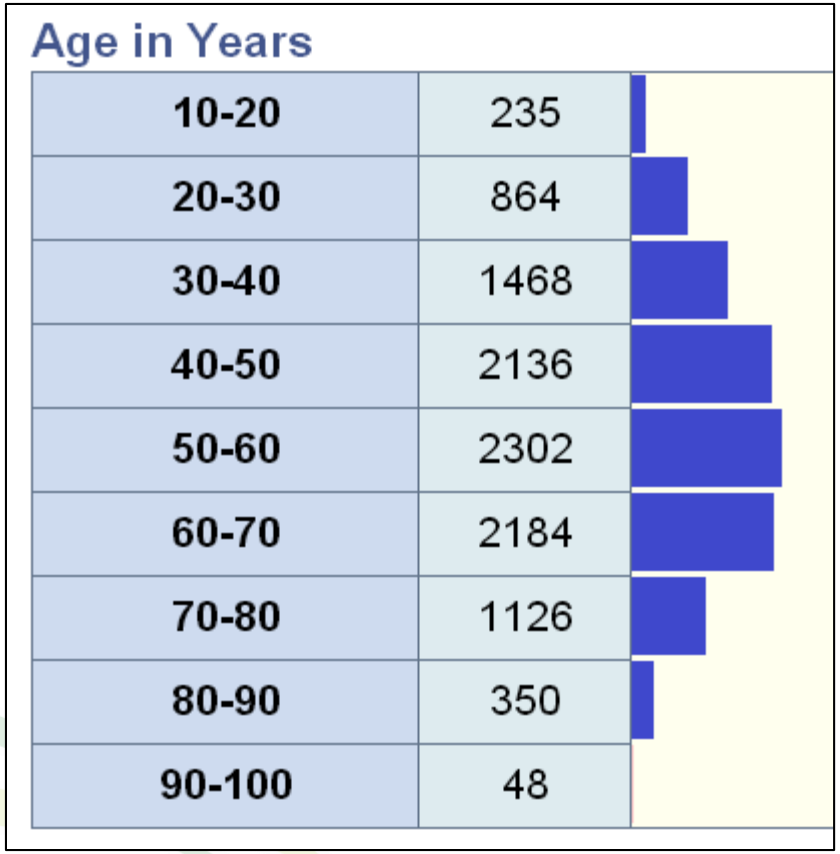
High blood pressure

Depression

Obesity

Osteoarthritis

Diabetes



[1]Self-reported data collected at baseline, n=10,562, June 2014. [2]Based on 104 Census tracts covered by MURDOCK catchment region.

MURDOCK STUDY

Lifestyle Habits: Diet, Past Year

Servings, fruit & veggies (#)	
1	1%
2	21%
3	27%
4	17%
5	10%
6	5%
6-10	3%
11-20	0%
21-40	0%
Don't know	14%

Sugar sweetened beverages (#)	
1	46%
2	23%
3	11%
4	5%
5	2%
6	1%
6-10	1%
11-20	0%
21-40	0%
Don't know	9%

Read nutrition labels	
Always	16%
Very often	30%
Sometimes	29%
Rarely	13%
Never	10%

N=9476 MURDOCK Study Participants

MURDOCK STUDY

Socioeconomic, lifestyle examples

Current Employment Status	N
NULL	0
Other	20
Permanently disabled	126
Retired	204
Stay at home full-time for parenting, care giving, or other responsibilities	11
Student	1
Temporarily laid off or sick/maternity leave	5
Unemployed/looking for work	32
Working now full-time	50
Working now part-time	45

Highest Education Level	N
Bachelor's degree	46
High school graduate (includes equivalent such as GED)	152
Less than high school graduate	57
Master's or higher professional degree	26
Some college or associate's degree	218

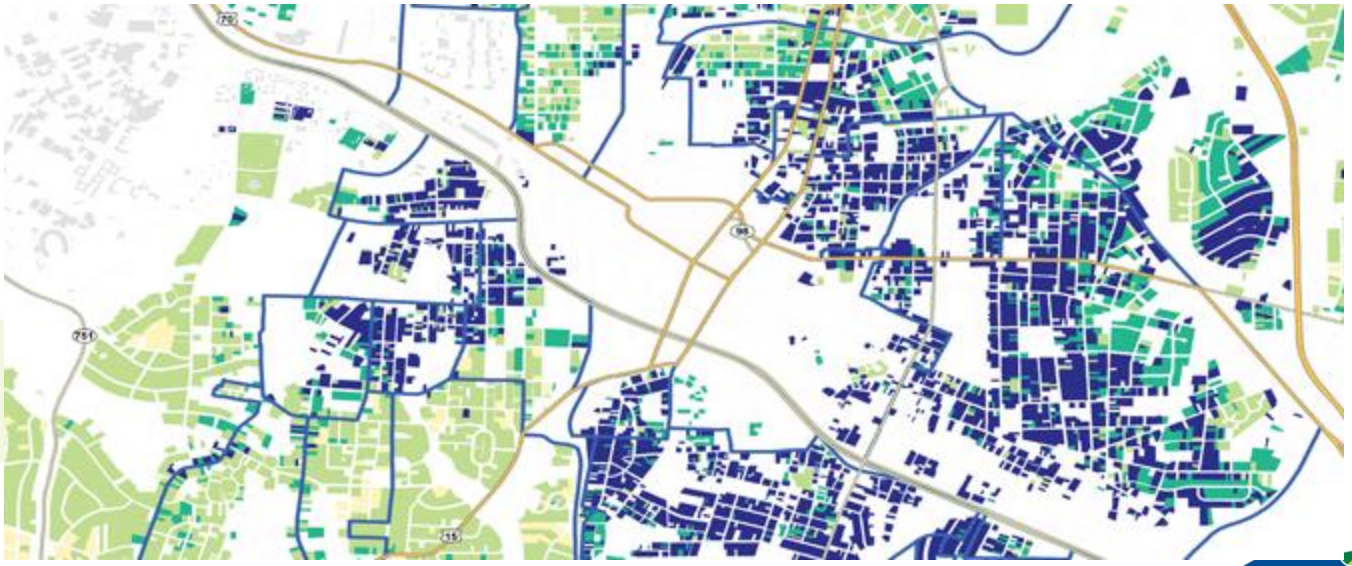
Total Household Income	N
NULL	0
$10,000-29,999	165
$30,000-49,999	95
$50,000-69,999	58
$70,000-89,999	33
$90,000 or more	39
Dont Know	46
Under $10,000	51

Currently Smoke Cigarettes	N
Yes, only some days	52
NULL	0
Yes, everyday	99
No	335

MURDOCK

Geospatial data

- Expand upon work done in Durham, NC by Dr. Marie Lynn Miranda (now at U Michigan)
- Know address- pinpoint where subjects live
- Include environment in outcomes analysis



Informatics challenge: free text medication data

- Map free text to coded terminology to enable querying by medication
  - Query by drug name (handle variations and misspellings)
  - Query by category (e.g. all SSRIs)
  - i2b2 as front end

Standard terminology #1: RxNorm

- Produced by NLM
- 11 source terminologies
- Normalized naming system for generic and branded drugs
- Tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems
- API accessible

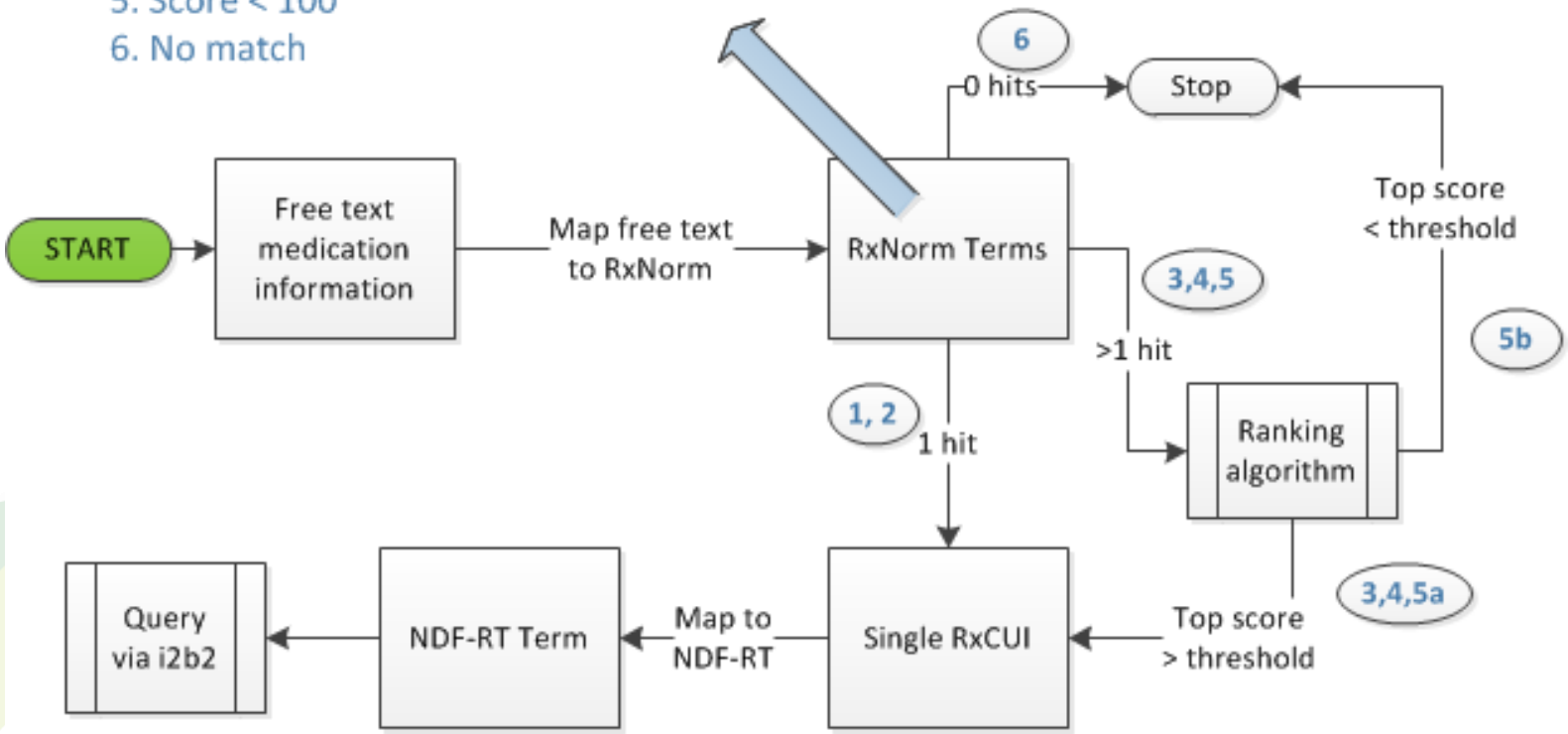
Standard terminology #: NDF-RT

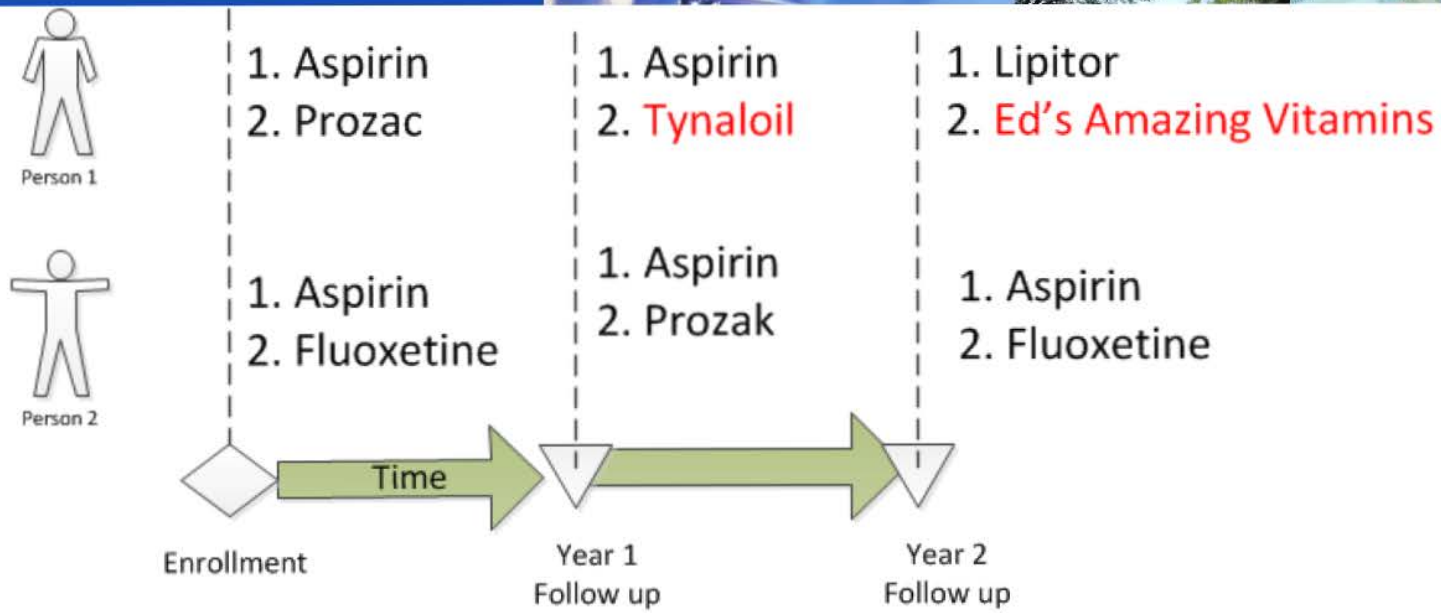
- National Drug File Reference Terminology
- Extension of VHA National Drug File (NDF)
- Produced by VHA (Veterans Health Administration)
- Organizes the drug list into a formal representation
- Used for modeling drug characteristics, e.g. drug class– beta blockers, anti-depressants, etc.
- Updated monthly

Text→ RxNorm mapping workflow

5 outcomes:

1. Perfect match- no score given
2. Exactly 1 hit with score of 100
3. N hits with score of 100, top hit is non-proprietary
4. Proprietary-only hits with score of 100
5. Score < 100
6. No match





Total:

- Aspirin- RXCUI 1191
- Aspirin- RXCUI 1191
- Aspirin- RXCUI 1191
- Aspirin- RXCUI 1191
- Aspirin- RXCUI 1191
- Lipitor- RXCUI 153165
- Prozac- RXCUI 58827
- Prozak- RXCUI 58827
- Fluoxetine- RXCUI 4493
- Fluoxetine- RXCUI 4493
- Ed's Amazing Vitamins- ???
- Tynaloil- RXCUI 584198 (Tygacil)

10/12 mapped correctly
1 mapped incorrectly
1 not mapped

Unique:

- Aspirin- RXCUI 1191
- Prozac- RXCUI 58827
- Prozak- RXCUI 58827
- Fluoxetine- RXCUI 4493
- Lipitor- RXCUI 153165
- Ed's Amazing Vitamins- ???
- Tynaloil- RXCUI 584198 (Tygacil)

5/7 mapped correctly
1 mapped incorrectly
1 not mapped

Accuracy

At the given threshold:

Q: What % of total entries did we code correctly?
A: 83% total terms (10/12)

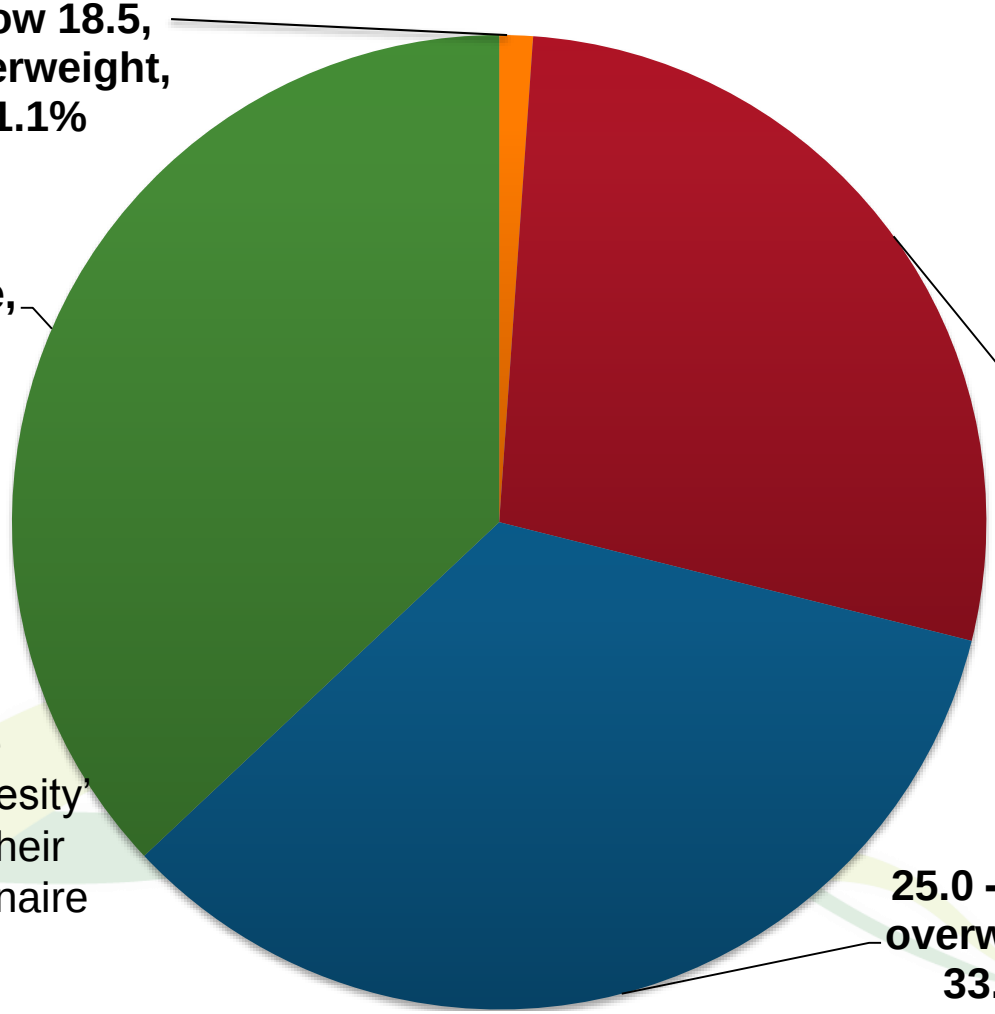
Q: What % of unique entries did we code correctly?
A: 71% unique terms (5/7)

Q: What % of people did we get perfectly correct?
A: 50% of participants (1/2)

Conclusions

- Given threshold of score ≥ 50:
  - Correctly mapped ~94% (122,523) of total terms
  - 5% unmapped (6,510)
  - "false positive rate" (i.e. mapped, but incorrectly) of ~1% (1,240 entries)
- Poor mapping happens largely with OTC, e.g. vitamins, minerals, supplements
- Important to consider use cases in deciding where to set cutoff
- RxNorm and APIs are updated periodically. Caveat emptor, and sign up for alerts.

Calculated BMI of MURDOCK Cohort



Only 26.7% of the cohort reports 'obesity' as a condition in their baseline questionnaire

EHR Data Quality

- The consistency and completeness of the MURDOCK Self-report data is well characterized and regularly audited.
- The accuracy of 1) the self-report data and 2) the EHR data have not been assessed
  - We do not know what type of data discrepancies there are, what data elements are impacted, or how the data discrepancies are otherwise distributed
  - Need to know PPV and NPV
- Availability and completeness of the EHR data have not been assessed
- These should be assessed to support continuing use of the data.

PI: Meredith Nahm Zozus, PhD

EHR Data Requested

Table 3: Summary of EHR Clinical Data Domains to be Requested from Healthcare Providers

Structured data	**Narrative Text data**
Demographics	Admission Reports
Problem List and claims diagnosis & Procedure codes	Procedure Reports
Labs (inpatient, ED and outpatient)	Diagnostic test reports
Vital signs (inpatient, ED and outpatient)	Clinical Notes
Medications (orders and med-reconciliation history file)*	Discharge Summaries
Inpatient Encounters (date, procedures and diagnoses)	Referral and consult notes
Outpatient Encounters (date, procedures and diagnoses)	
Emergency Department Visits (date, procedures, and diagnoses)	
Smoking Status	
Medical History	

*will confirm with facilities whether medications will come with an NDC code.

MURDOCK STUDY

Informatics challenge: record linkage optimization

- Collaborate with experts at Univ. Arkansas Little Rock to optimize linkage algorithms

Preliminary Results, replication & reconciliation in progress

	First Name	Middle Name	Last Name	Date Of Birth	Address	City	Sex
Rule1	Scan		Scan	Scan(first 5 digit)			
Rule2	Scan		LED(0.5)	Scan			
Rule3	Scan		Scan		Scan		
Rule4			Scan	Scan		Scan	Scan
Rule5	Soundex			Scan	Scan	Scan	

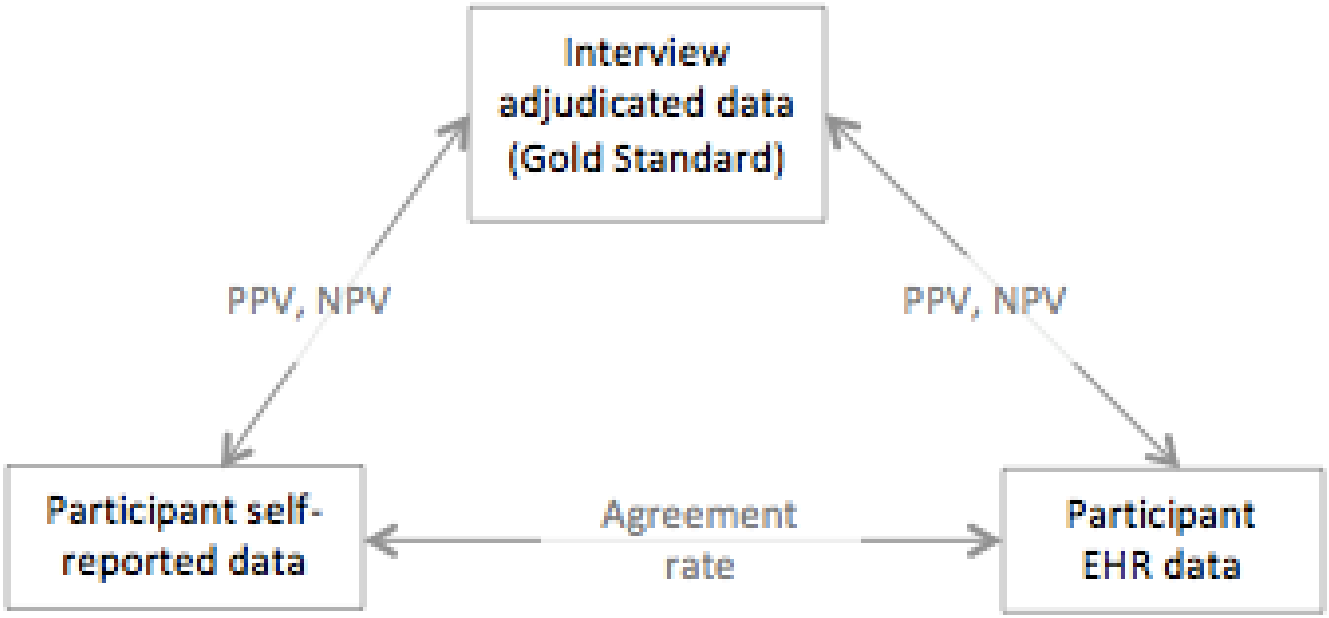
*From Wang & Talburt presentation at team meeting May 27, 2014

3 False Negative Records

Last Name	F_Name	DOB	Address	City	State	Zip-5
Martinez	Mirabel	04/03/80	323 McCormick Dr	Concord	NC	28027
Martinez, de la	Mirabel	04/03/80	2930 Hidden Field Dr	Concord	NC	28027
Kliman	Judy	08/19/64	2023 N Earnheart Blvd	Kannapolis	NC	28083
Kliman	Judy	08/15/64	208 Morrison Ct	Concord	NC	28027
Ware	Roberta	11/05/80	3009 Dog Tree Ln	Charlotte	NC	28213
Ware	Roberta	11/18/80	200 St Albans Ave	Concord	NC	28027

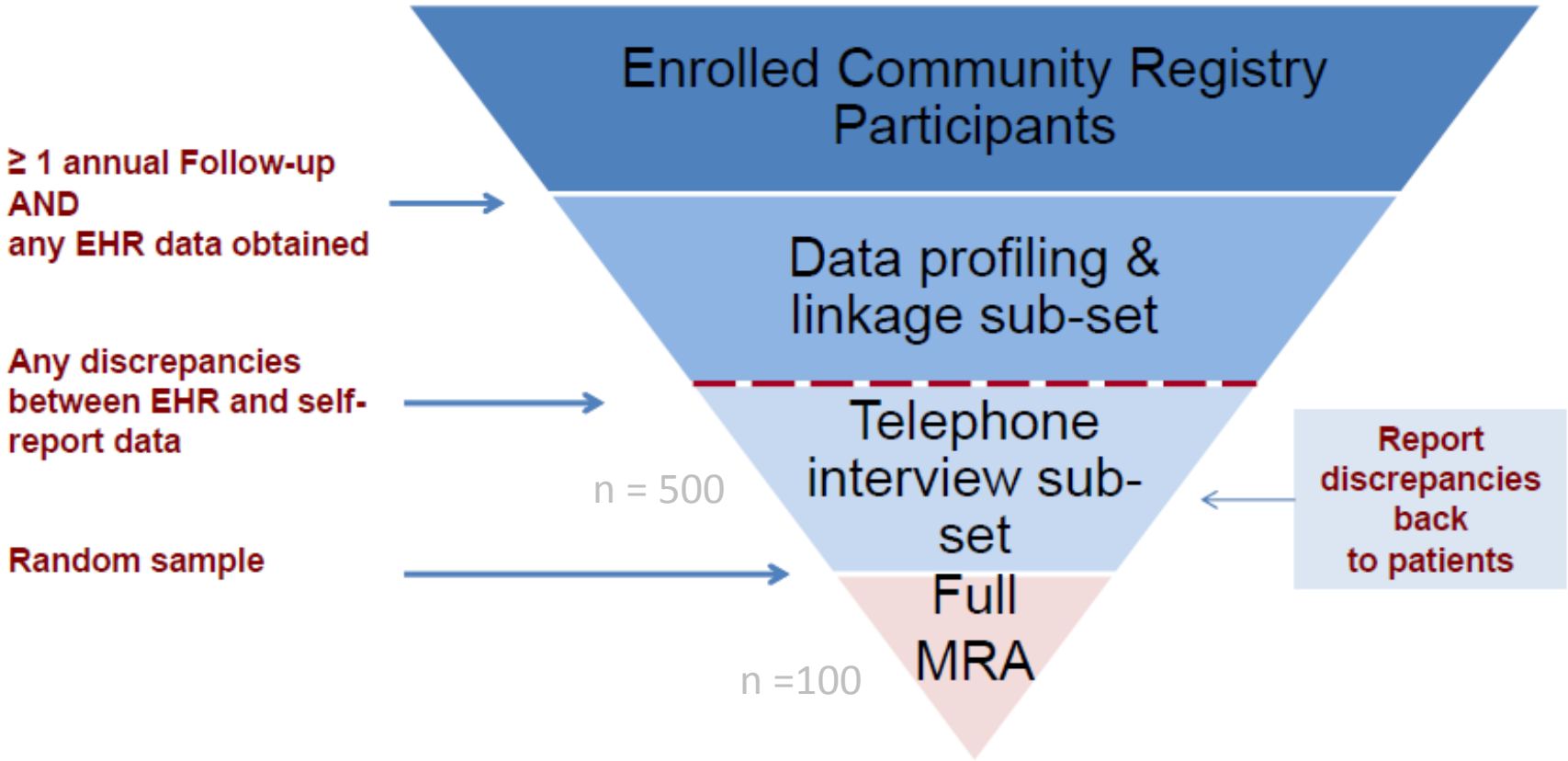
Information Changed with similar variation to show example False Negatives

Back to data quality…

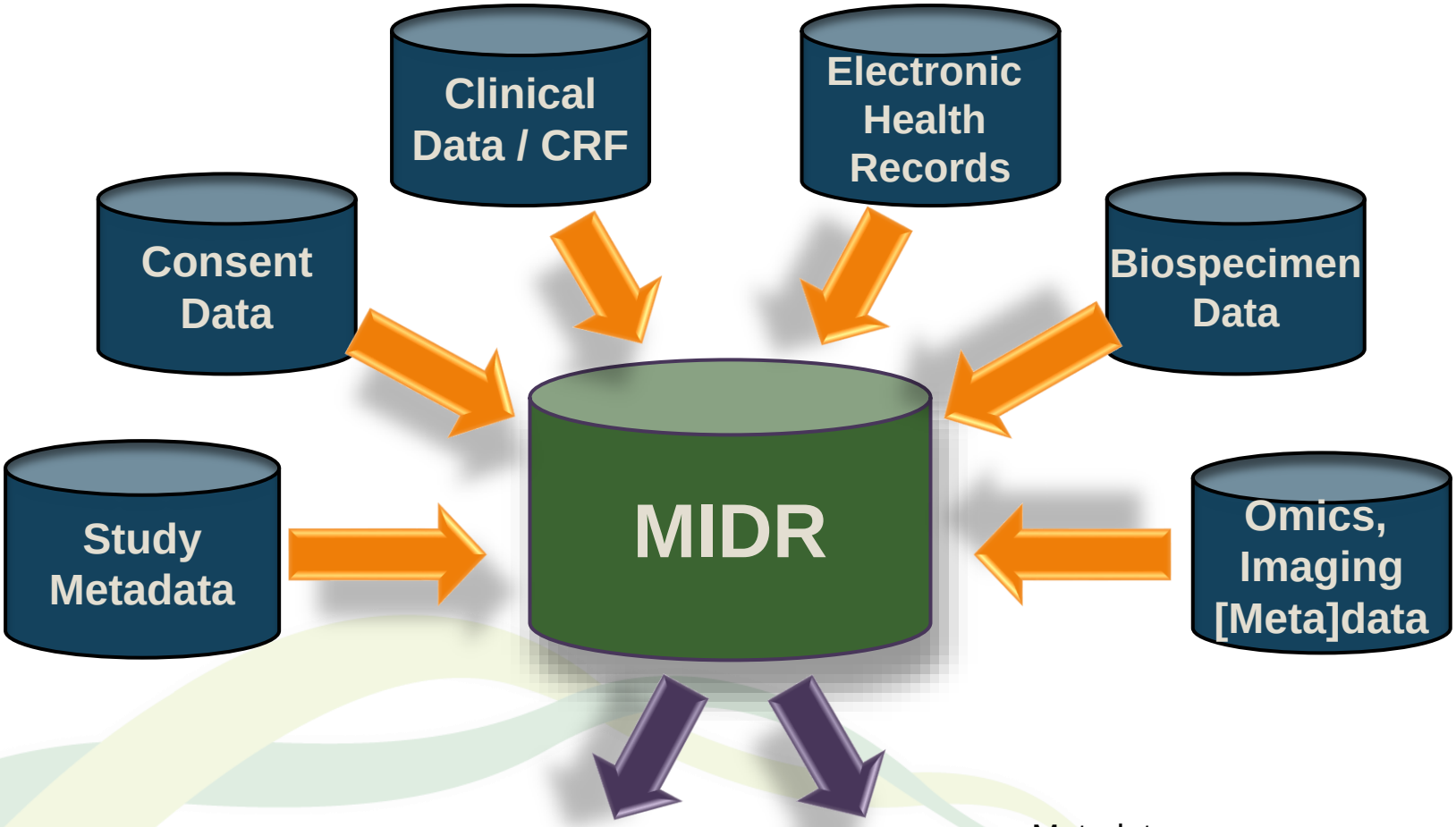


Study Structure

%



The MURDOCK Integrated Data Repository (MIDR)



Data mining, disease stratification, predictive algorithms, pathway modeling, systems biology

Discovery

Information Retrieval

Metadata queries, cohort selection, administrative reports

MURDOCK STUDY

MURDOCK Integrated Data Repository (MIDR)

- Relational Registry database imported to i2b2
- Validated and deployed to MURDOCK team in January 2014

- Current:
  - Participant questionnaire info (demographics, lifestyle, medical history, medications)
  - Biospecimen data
  - Omic metadata
  - Study metadata
- Future:
  - Up-to-date biospecimen data
  - Omic data (likely using tranSMART)
  - Structured consent data
  - EHR data

MURDOCK STUDY

i2b2 demo

i2b2 Query & Analysis Tool | Project: H1.5 Repository Test 06-28-2014 | User: Melissa Cornish | Find Patients | Analysis Tools | Message Log | Help | Logout

Navigate Terms | Find Terms

- Biomarker Data
- Biospecimens
- Clinical
- Demographics
- Lifestyle
- Study Metadata
- Visit

Workplace

- mc174

Previous Queries

- Gender@07:03:04 [6-12-2014] [mc174]
- Gender@06:37:35 [6-12-2014] [mc174]
- On Average Day @22:16:23 [6-11-2014] [mc174]
- 21-25 Years@22:14:22 [6-11-2014] [mc174]
- How Many Years @22:13:41 [6-11-2014] [mc174]
- Currently Smoke@22:12:39 [6-11-2014] [mc174]

Query Tool

Query Name:

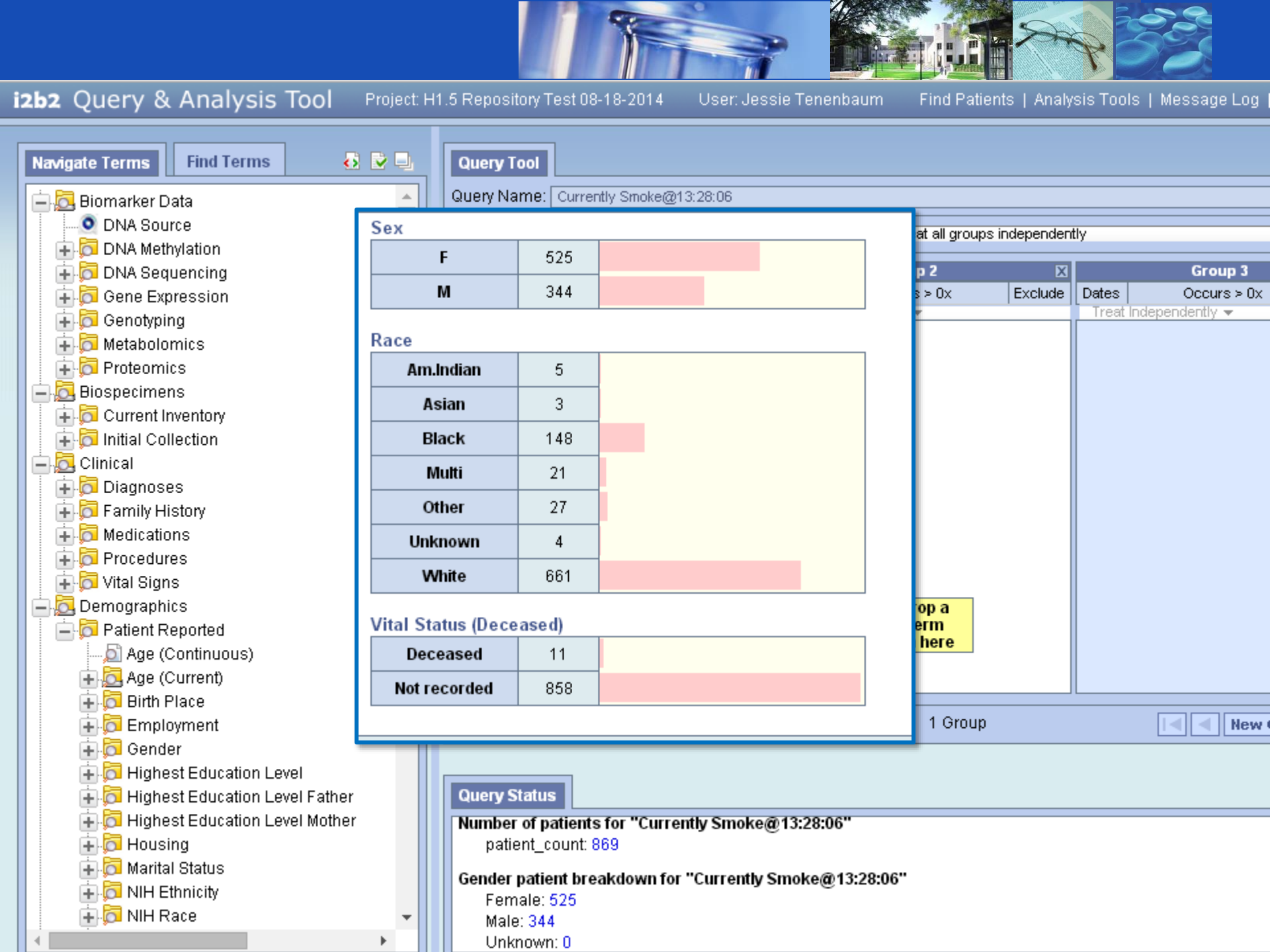
Temporal Constraint: Treat all groups independently

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		

drop a term on here

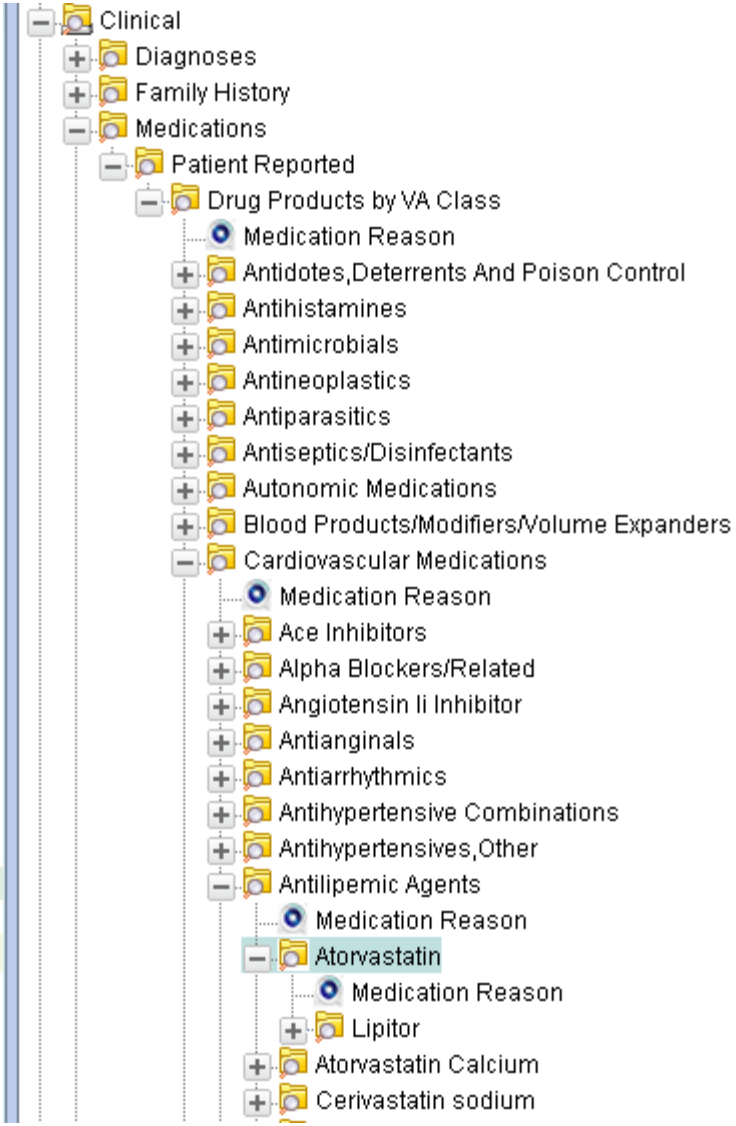
Run Query | Clear | Print Query | 0 Groups | New Group

Query Status


i2b2 Query & Analysis Tool
Project: H1.5 Repository Test 08-18-2014
User: Jessie Tenenbaum
Find Patients | Analysis Tools | Message Log
Navigate Terms
Find Terms
Biomarker Data
DNA Source
DNA Methylation
DNA Sequencing
Gene Expression
Genotyping
Metabolomics
Proteomics
Biospecimens
Current Inventory
Initial Collection
Clinical
Diagnoses
Family History
Medications
Procedures
Vital Signs
Demographics
Patient Reported
Age (Continuous)
Age (Current)
Birth Place
Employment
Gender
Highest Education Level
Highest Education Level Father
Highest Education Level Mother
Housing
Marital Status
NIH Ethnicity
NIH Race
Query Tool
Query Name: Currently Smoke@13:28:06
Sex
F 525
M 344
Race
Am.Indian 5
Asian 3
Black 148
Multi 21
Other 27
Unknown 4
White 661
Vital Status (Deceased)
Deceased 11
Not recorded 858
Group 3
Dates
Occurs > 0x
Treat Independently
1 Group
Query Status
Number of patients for "Currently Smoke@13:28:06"
patient_count: 869
Gender patient breakdown for "Currently Smoke@13:28:06"
Female: 525
Male: 344
Unknown: 0

Medications, e.g.

- ACE Inhibitors
- Lipitor
- Antiparkinson agents



MURDOCK

Navigate Terms | Find Terms

- Biomarker Data
- Biospecimens
  - Current Inventory
    - Containers
    - Notes
    - Blood Specimen
      - Containers
      - Fasting
      - Notes
      - Pregnant
      - Buffy Coat
      - Paxgene
      - Plasma
      - Serum
      - Whole Blood
        - Containers
        - Fasting
        - Notes
        - Pregnant
    - Derivatization
    - Urine
  - Initial Collection
- Clinical
  - Diagnoses
    - Patient Reported
      - Alzheimers Disease
      - Asthma
      - Atrial Fibrillation
      - Breast Cancer
      - Cervical Cancer
      - Colon Cancer
      - Congestive Heart Failure
      - Coronary Artery Disease
        - Coronary Artery Disease Baseli
        - Coronary Artery Disease Follow
      - Crohns Disease

Query Tool

Query Name: Coron-Coron-Whole@15:21:51

Temporal Constraint: Treat all groups independently

Group 1	Group 2	Group 3
Dates / Occurs > 0x / Exclude	Dates / Occurs > 0x / Exclude	Dates / Occurs > 0x / Exclude
Treat Independently	Treat Independently	Treat Independently
y Artery Disease Baseline = ("No")	Artery Disease Followup = ("Yes")	Whole Blood >= 1 vials
one or more of these	one or more of these	one or more of these

AND AND

Run Query | Clear | Print Query | 3 Groups | New Group

Query Status

Finished Query: "Coron-Coron-Whole@15:21:51" [5.0 secs]

Compute Time: 4 secs

Number of patients for "Coron-Coron-Whole@15:21:51"

patient_count: 95

Focused Cohorts Building on Established Infrastructure (Horizon 2)

Cohort	Data Collected in Addition to Samples
Memory & Cognitive Health Screener	>1500 over age 55 screened using Montreal Cognitive Assessment (MoCA), cognitive screener (ADCS), word list memory, Reitan trail making, and delay word memory tests *Re-screening coming soon*
Multiple Sclerosis	>500 diagnosed with MS screened using disease specific questionnaire to explore open platform proteomics in an exquisitely well-phenotyped MS population (n=1000)
Primary Progressive Multiple Sclerosis	>7 currently followed with bi-annual sample collection; sample not available elsewhere in the US or world (n=100)
Physical Performance and Healthy Aging	>600 screened with memory and physical assessments to detect genetic, physical and behavior changes that influence cognitive and physical development. Data and samples collected at baseline and 2 year follow-up visit (n=1000)
Severe Acne	>120 severe acne patients who have undergone or are undergoing standard Isotretinoin treatment (n=250)

MURDOCK STUDY

Acknowledgements

MURDOCK Leadership

- Rob Califf
- Victoria Christian
- Geoff Ginsburg
- Ashley Dunham
- L Kristin Newby
- Meredith Nahm
- Krishna Udayakumar
- Melissa Cornish

Kannapolis Team, esp:

- Perla Nunes
- Leah Bouk
- Sarah Maichle

Also

- Doug Wixted
- Brian Bennett

Informatics

- Colette Blach
- Anita Walden
- Julie Frund
- Kurt Morehouse
- Maryam Younes
- David Steele
- Michelle Smerek
- Guilherme Del Fiol

And the rest of the team at www.murdock-study.com

Funding: NCRR grant UL1RR024128 (Duke CTSA), **David H Murdock**

MURDOCK STUDY

Questions?

Jessie Tenenbaum, PhD
Associate Director, Bioinformatics
+1.919.668.8811
jessie.tenenbaum@duke.edu

Melissa A. Cornish, MSPH
Business Development
+1.919.622.2254
melissa.cornish@duke.edu

Visit our website: www.murdock-study.org