THE UNIVERSITY OF
CHICAGO
MEDICINE &
BIOLOGICAL
SCIENCES

CENTER FOR RESEARCH INFORMATICS
SEMINAR SERIES

FREE TRAINING SESSIONS FOR THE UCHICAGO COMMUNITY

Optimizing Data Requests for the Clinical Research Data Warehouse

Tim Holper, MS, MA
Director of Data Warehouse & Business Intelligence, CRI

Julie Johnson, MPH, RN
Senior Healthcare Business Analyst, CRI

November 3, 2016

Disclaimer and Permissions

- This educational activity is being presented without the provision of commercial support and without bias or conflict of interest from the planners and presenters.

- The purpose of these materials is to provide insight into the variety of data sources, as well as the recommended process for obtaining data from the Analytics Core within UCMC

- These materials are not meant to provide an exhaustive analysis of all data sources and means for requesting data within the institution

- It is not permissible to share these materials outside UCMC

Agenda

- What is a Data Warehouse?
- What is the CRDW?
- CRDW I/O
- Data Warehousing Challenges
- Quality of Data Sources
- Other Data Warehouses at UCMC
- How do I request data?
- Analytics Core and ACReS
- CRDW Request Triage
- Data Request Challenges
- Regulatory Compliance
- Semi-Self Service Tools
- Questions?

THE UNIVERSITY OF
CHICAGO MEDICINE &
BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

What Is a Data Warehouse?

THE UNIVERSITY OF
CHICAGO MEDICINE &
BIOLOGICAL SCIENCES

CENTER FOR
RESEARCH
INFORMATICS

What Is a Data Warehouse?

What Is a Data Warehouse?

- Bill Inmon, generally considered to be the father of modern data warehousing, describes a data warehouse as:
  '**A subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions**'

- What does this mean?

- A Data Warehouse is a repository of very specific and very static data organized in ways (usually by time and by some other set of attributes), updated on a regular basis, so that the data can be used by folks to make decisions
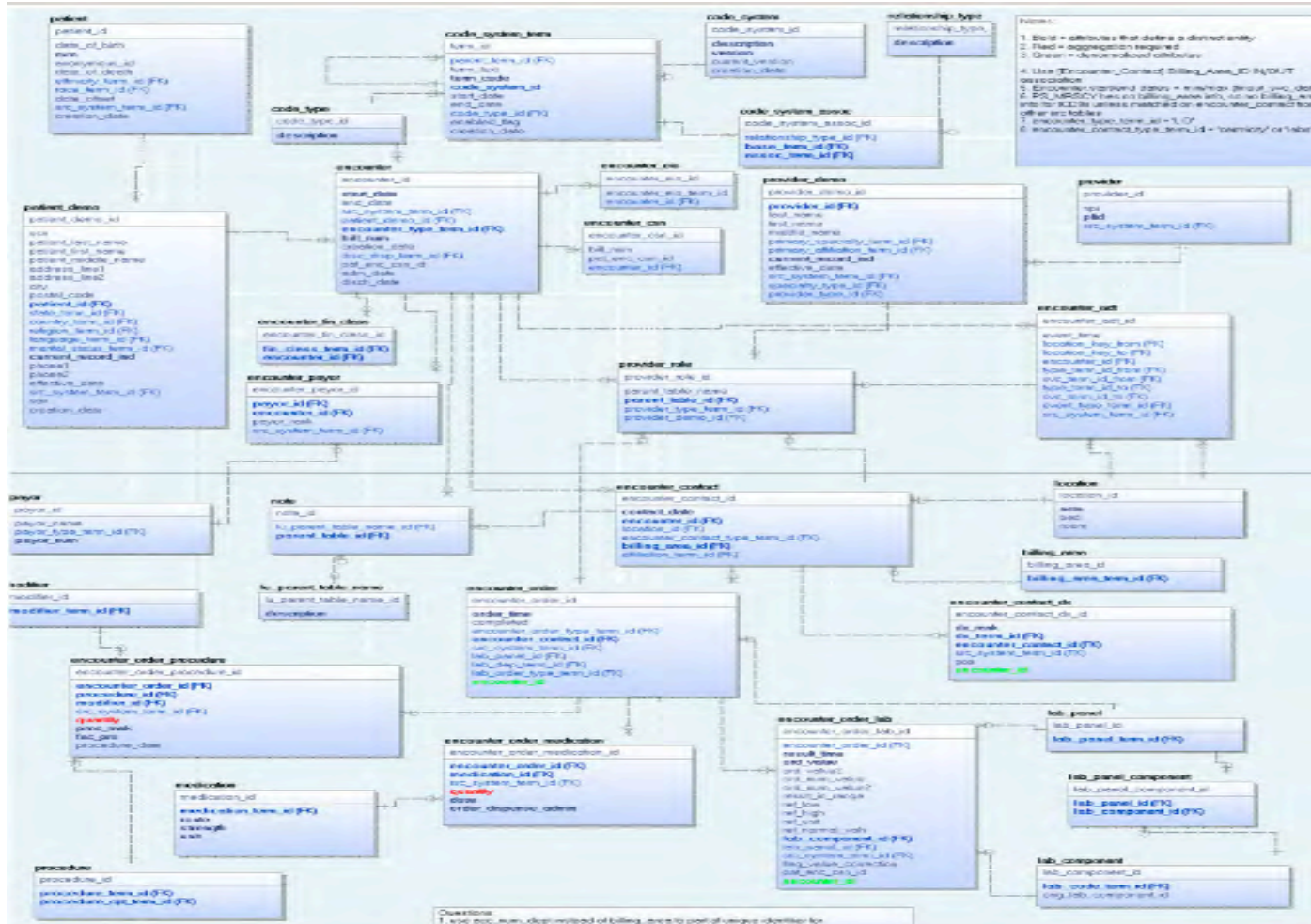
What Is a Data Warehouse?

- **Subject Oriented:** Data organized into logical subject areas

- **Integrated**: Removal of inconsistencies regarding naming convention and value representations

- **Nonvolatile**: Data stored in read-only format and updated daily

- **Time Variant:** Data not in real-time

What Is a Data Warehouse?

- For the purposes of this discussion, a Data Warehouse is a central repository for storing and extracting the **most commonly-requested data elements** required to answer a question

THE UNIVERSITY OF
CHICAGO MEDICINE &
BIOLOGICAL SCIENCES | CENTER FOR
RESEARCH
INFORMATICS

Data Warehouse Is a Physical Database

Data Warehousing Is a Process

More than just hardware and software

The fastest hardware and the latest and greatest software is only a **small part of that process**

Without good processes in place for managing data, requesting data, governance of data, consistent standards, and quality control, the best hardware and software will do one thing: **serve up bad data faster**

What Is the CRDW?

- CRDW is a repository of University of Chicago medical data dating back to 2006. The CRDW team brings together data from disparate sources, including Epic electronic medical records, Centricity billing, Cancer Registry, REDCap, and Labvantage to create cohesive datasets for research.

- CRDW is an interface for 'seamlessly' integrating disparate data sources

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

What Is IN the CRDW?

- The **most commonly-requested data elements** required to answer research questions are the following:
- Patient Demographic Info
- Encounter Info (Encounter type, Admit Date, Discharge Date, etc.)
- Diagnoses: ICD9/10
- Procedures: ICD9/10, CPT
- Flow Sheets: Vitals, Respiratory, Physical Assessment, etc.
- Medications: Outpatient and MAR
- Labs
- ADT: (Admission, Discharge, Transfer)

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

What About Notes?

- Notes are available in bulk and we can provide as part of a data request

- However, defining a cohort or doing additional filtering/logic using info from within the note is difficult, time-consuming, and expensive using Natural Language Processing

- Not impossible, just cost-prohibitive
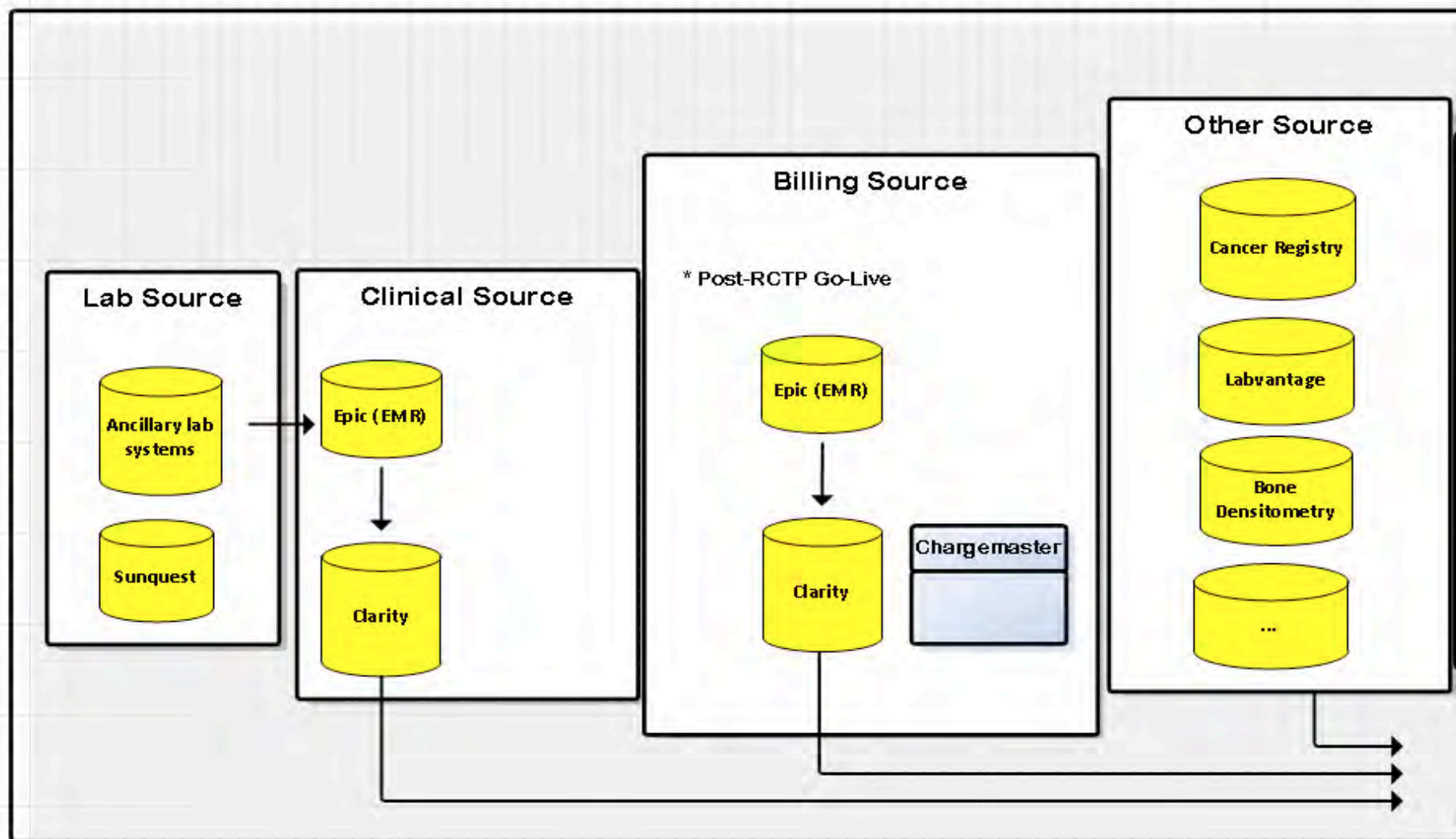
Source Systems- Internal to UCMC

- UCM source data lives within a number of disparate source systems around the University of Chicago Medical Center

- At a sufficiently high level, data sources break down into the following areas:
  - Clinical
  - Billing
  - Other

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

Source Systems- Internal to UCMC
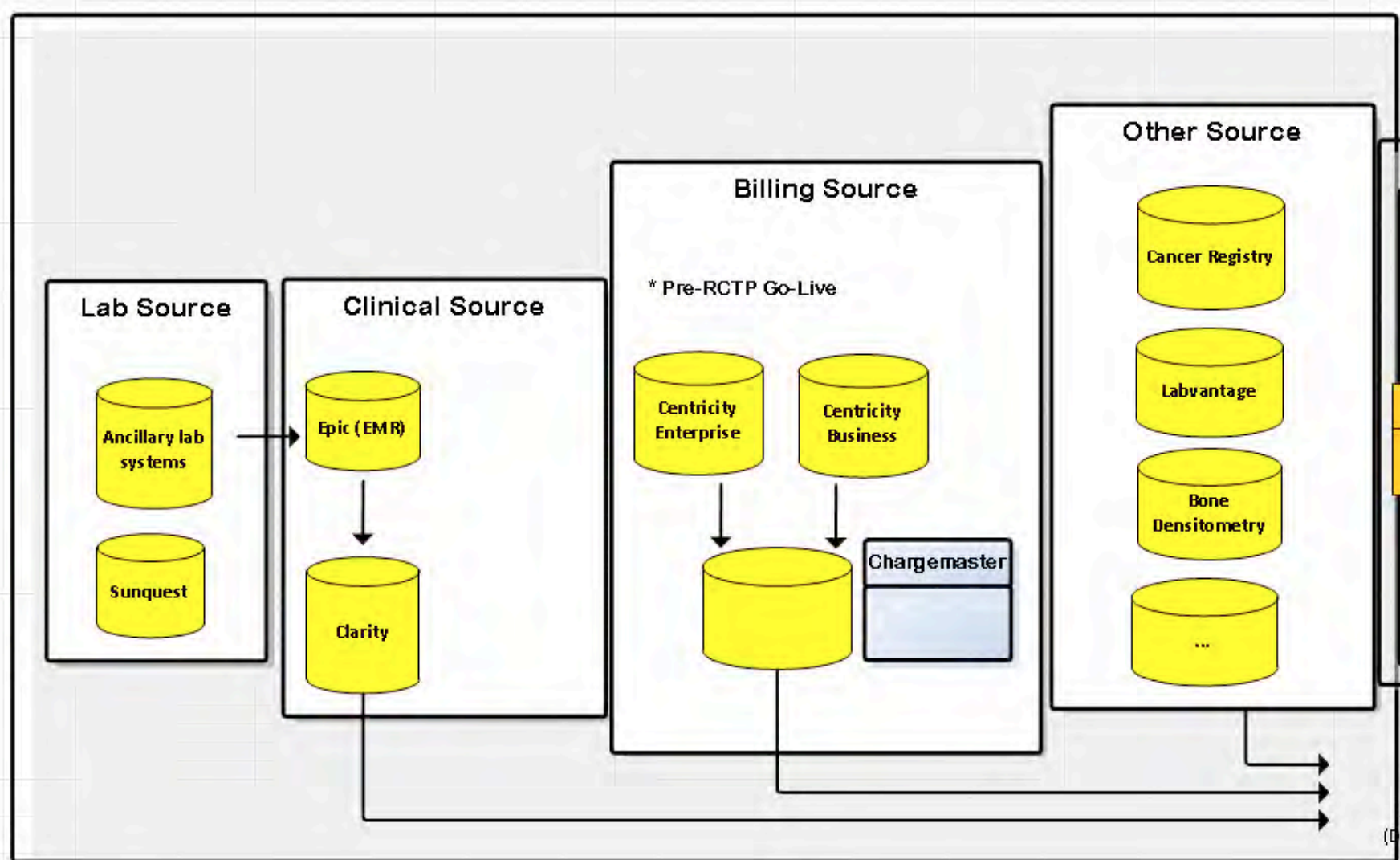
Source Systems- Internal to UCMC

Source Systems- External to UCMC

- External 'Publicly Available' Data Sources, which are typically population-based, rather than specific to UCMC

- Examples include the following:
  - Medicare/Medicaid Data - MarketScan
  - SSA's National Death Registry

- In most cases, we can point you in the right direction

- Challenges of matching data from outside of the institution

- Example: National Death Registry:
  - On a monthly cadence, we pull SSA's National Death Registry (NDR) file into the CRDW and match records w/ patients in Epic
  - While this can only be a probabilistic match, using a combination of SSN, DOB, Last Name, and First Initial, we see from 110k - 150k+ patients in Epic which lack accurate mortality data.

THE UNIVERSITY OF
CHICAGO MEDICINE &
BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

Output from CRDW: 2 Flavors

- Static Data Set: 1x data extraction
  - Example: Finding a cohort of patients that match on a specific set of demographic, clinical, and billing variables.

- Dynamic Data Set (AKA Affiliated Data Mart)
  - Example: Bringing together a registry of patients maintained in REDCap w/ clinical data from Epic/Clarity, billing data from Centricity, and cancer data from the Cancer Registry

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

Output from CRDW

Table 1: COHORT

patient_id	race	ethnic_group	sex	dob_off	marital_status	bmi
123456	Black/African-American	Not Hispanic or Latino	F	10/31/1940	Married	24.03
234567	Black/African-American	Not Hispanic or Latino	F	12/25/1951	Single	39.89
345678	White	Not Hispanic or Latino	F	3/12/1937	Married	23.24
456789	Black/African-American	Not Hispanic or Latino	F	2/6/1969	Single	47.54
567890	Black/African-American	Not Hispanic or Latino	F	9/1/1935	Widowed	37.59

Table 2: PATIENT ENCOUNTERS

patient_id	encounter_id	enc_date_off	enc_type	fin_class	sec_payor
123456	901234567	10/20/2014	Outpatient	Medicare	Commercial
234567	890123456	9/2/2014	Outpatient	Medicaid	NULL
345678	789012345	7/10/2014	Inpatient	Commercial	Commercial
456789	678901234	1/13/2014	Outpatient	Commercial	NULL
567890	567890123	7/29/2013	Inpatient	Self-pay	NULL

Table 3: LABS

patient_id	encounter_id	order_time_off	proc_code	proc_name	result_time_off	ord_value
123456	901234567	12/23/2013	LABCHLDLP	LIPID PANEL	12/23/2013	78
234567	890123456	12/23/2013	LABENINSLE	INSULIN	12/26/2013	69.7
345678	789012345	12/23/2013	LABLBCBC3	CBC W/ DIFF & PLATELETS	12/23/2013	0.03

Table 4: DIAGNOSES

patient_id	encounter_id	enc_date_off	icd9_dx
123456	901234567	11/14/2012	496, 784.0,
234567	890123456	11/1/2014	250.00, 729.81, 787.20,
345678	789012345	8/23/2010	367.1, 367.4,
456789	678901234	12/14/2007	723.0, 724.5,
567890	567890123	9/4/2013	250.02, 272.2, 278.00, 401.1, V45.61,

Data Warehousing Challenges: Source Matters

- Some questions posed to the CRDW team, despite appearing straightforward on the surface, prove challenging to operationalize based on how we bring together disparate data sources.

- Example: Examining Patient intubation using clinical data vs. billing data

    – Patient is intubated and researcher wants to know the **clinical** date/time of intubation by **billing** code (i.e. CPT).
    – Using billing data, we have multiple service dates and can put the intubation procedure within a range of service dates, but we cannot necessarily pinpoint the exact timestamp for the intubation procedure based on billing code alone.
    – If we jump over to the clinical data in Epic, we have exact timestamps, but we cannot necessarily align those records w/ billing CPT codes.

- Take away: Parameters used to define a patient cohort can be very different from the data that actually shows up in the deliverable

- One way to think of it is in 2 steps:

    – Defining the cohort
    – Generating the data set

Data Warehousing Challenges: Crumpled-Up Cocktail Napkin

- Clients sometimes come to us w/ their own data

- Depending on how well the data has been curated, it can be difficult to align w/ CRDW data

- If you are collecting registry data locally, please consider the following suggestions:

1. Don't

2. Consider putting your data into REDCap

3. If you must/insist on storing data locally, please consider things like the following:

    - How data is stored (i.e. Dates in a 'date' field w/ appropriate data type in excel)
    - Limiting scope of individuals w/ permission to edit data
    - Additional fields which could provide more accuracy when eventually triangulating data within your registry w/ data from CRDW (i.e. billing / clinical encounter number).

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS
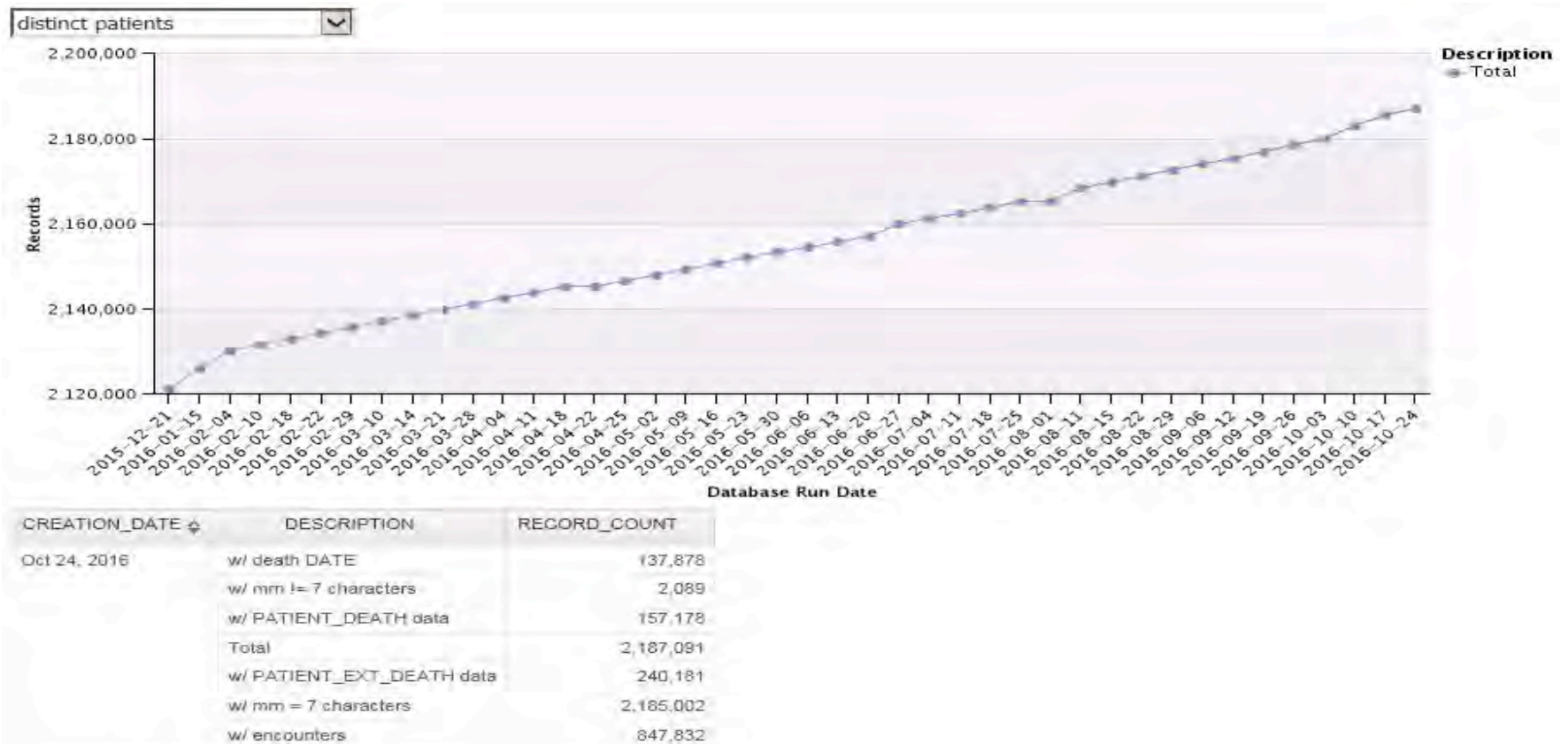
Data Warehousing Challenges: Discrete vs. Non-Discrete Data

- Researcher puts data into Epic in a discrete field

- Important to work w/ the right teams to make sure the data going into Epic is set up to eventually make its way over to Clarity where CRDW team takes over. More common w/ customizable fields (i.e. Smart Phrases, Doc Flow sheets)

- If discrete data is entered into discrete fields, but eventually makes its way over to Clarity within a blob of text, much more difficult for CRDW team to extract/parse

- Example: Researcher wants to define cohort based on extracting patient's Harvey Bradshaw Index from a progress note

  – Although Harvey Bradshaw Index is a discrete value, it may not be stored within a discrete field
  – Difficult to extract w/o NLP

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS
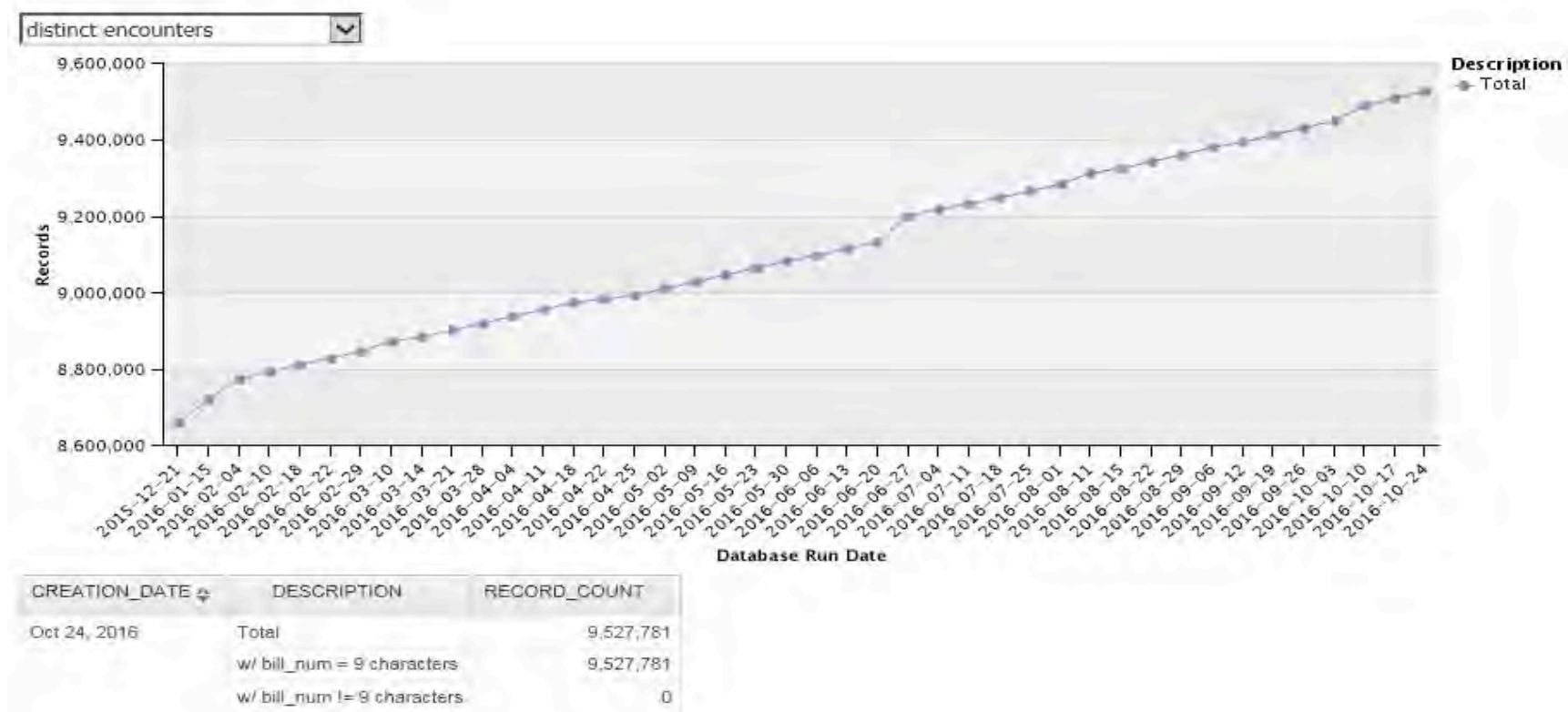
Quality of Data Sources

- Weekly reports follow trending in primary data sources

- Data should be stable and show a consistent pattern of variation over time

CREATION_DATE ⬦	DESCRIPTION	RECORD_COUNT
Oct 24, 2016	w/ death DATE	137,878
	w/ mm != 7 characters	2,089
	w/ PATIENT_DEATH data	157,178
	Total	2,187,091
	w/ PATIENT_EXT_DEATH data	240,181
	w/ mm = 7 characters	2,185,002
	w/ encounters	847,832

Quality of Data Sources

- Weekly reports follow trending in primary data sources

- Data should be stable and show a consistent pattern of variation over time

CREATION_DATE ⇱	DESCRIPTION	RECORD_COUNT
Oct 24, 2016	Total	9,527,781
	w/ bill_num = 9 characters	9,527,781
	w/ bill_num != 9 characters	0

Other Data Warehouses at UCMC

- **Nomise**: Facility Billing data (2012 – present)
  - Run by Managed Care

- **UCPG:** Professional Billing data (2012 – present)
  - Run by UCPG

Why So Many Data Warehouses?

- Different purposes

- If you have a request that is specific to quality or specific to research, the targeted repository is optimized for answering specific types of questions
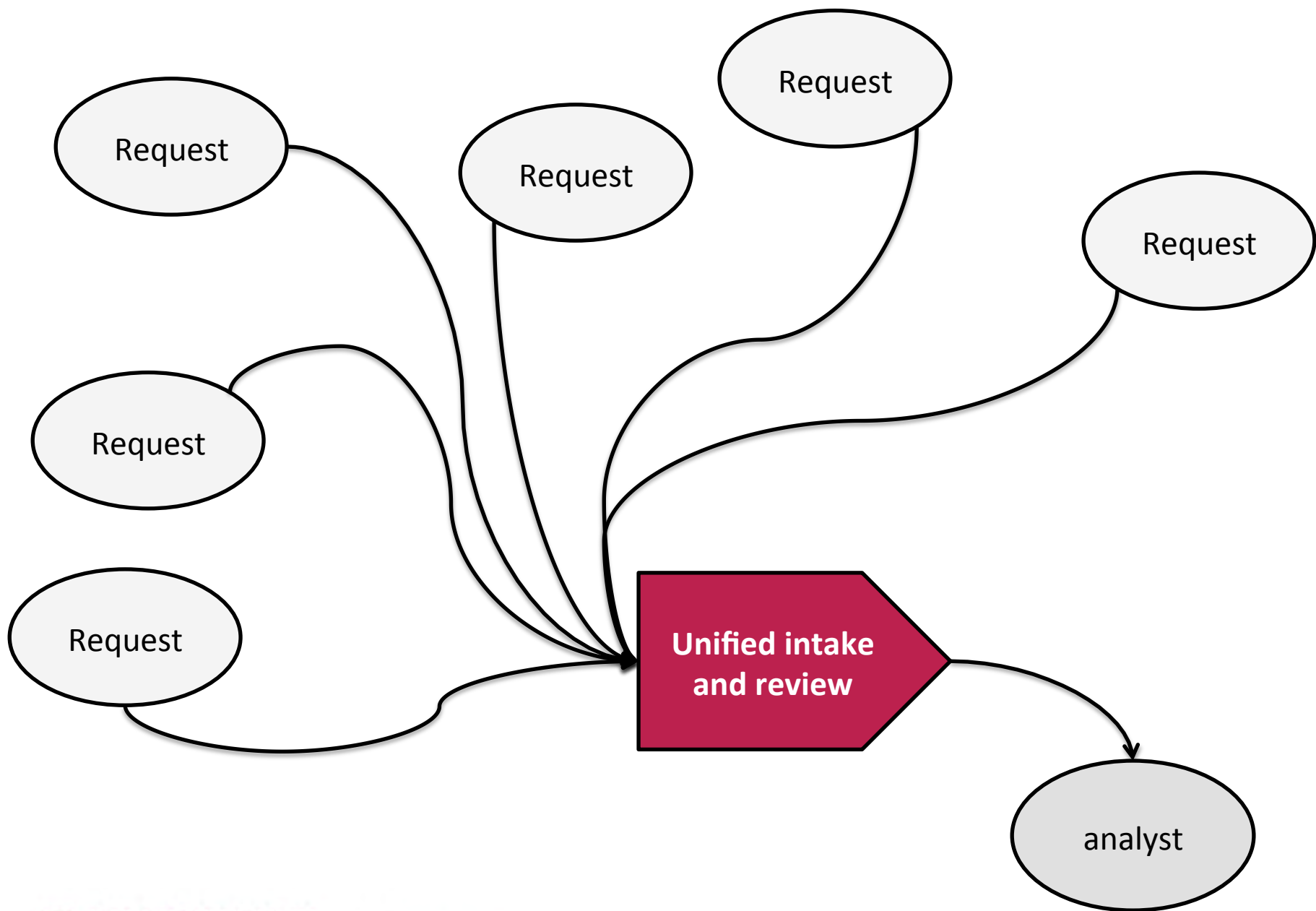
- Same data, different packaging

How do I request data?

History of Data Requests at UCM

- Not easy to obtain data when needed

  – Typically was driven by who you knew

- Inconsistency in way data was requested

  – Same requests filled by multiple teams

    • Discrepancy in outcomes
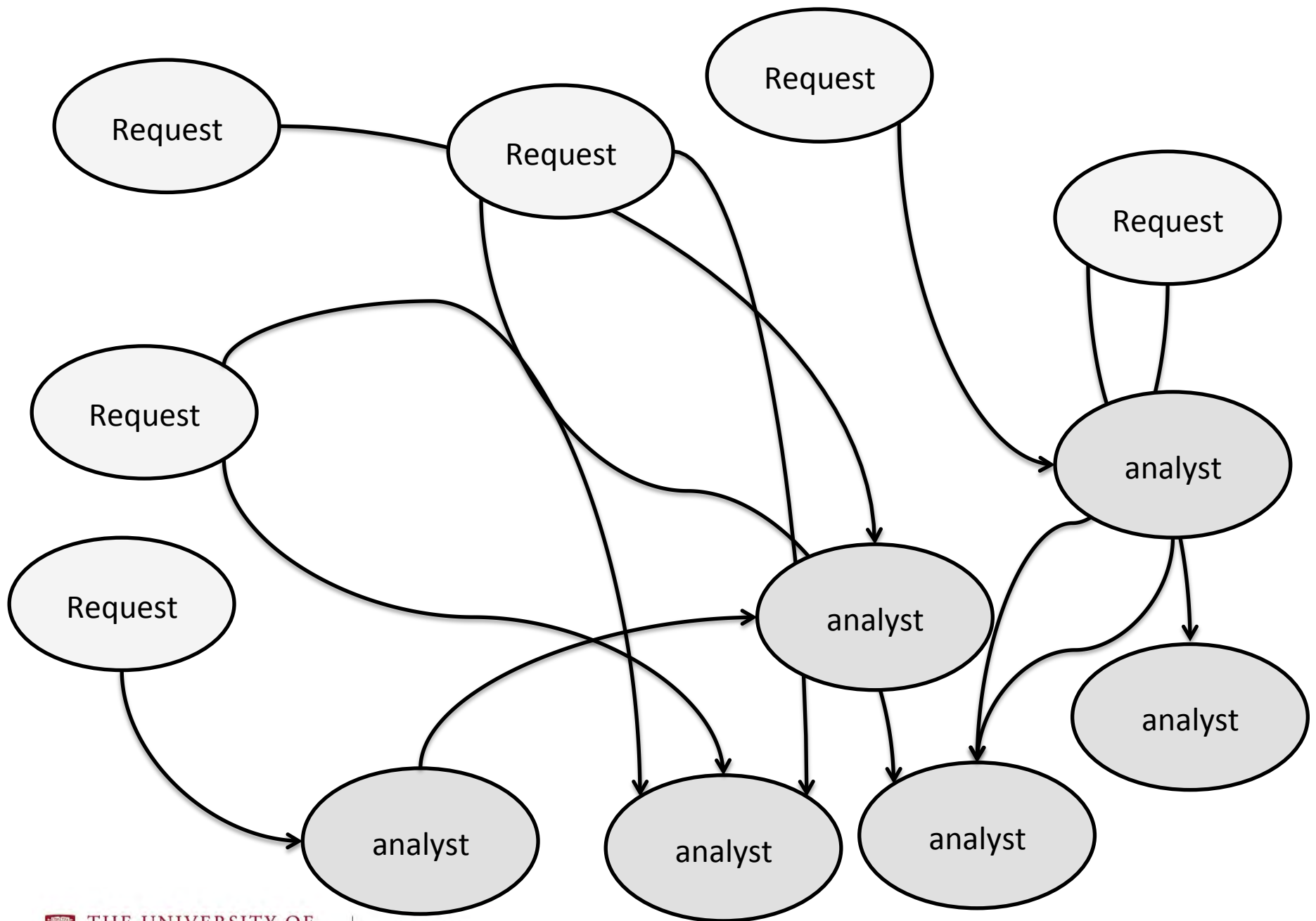
Analytics Core

- The Enterprise Informatics and Analytics Program leadership has chartered the Analytics Core to improve the path to data for the organization by fostering communication, collaboration, and best practices among the organization's analytics leaders

- Analytics Core is comprised of analytics leads from both technical and business teams spanning UCM and BSD

- Analytics Core meets weekly to review all data requests submitted via ACReS (Analytics Core Request System) to reduce duplication of resources, improve data quality, and provide a platform for communication and transparency

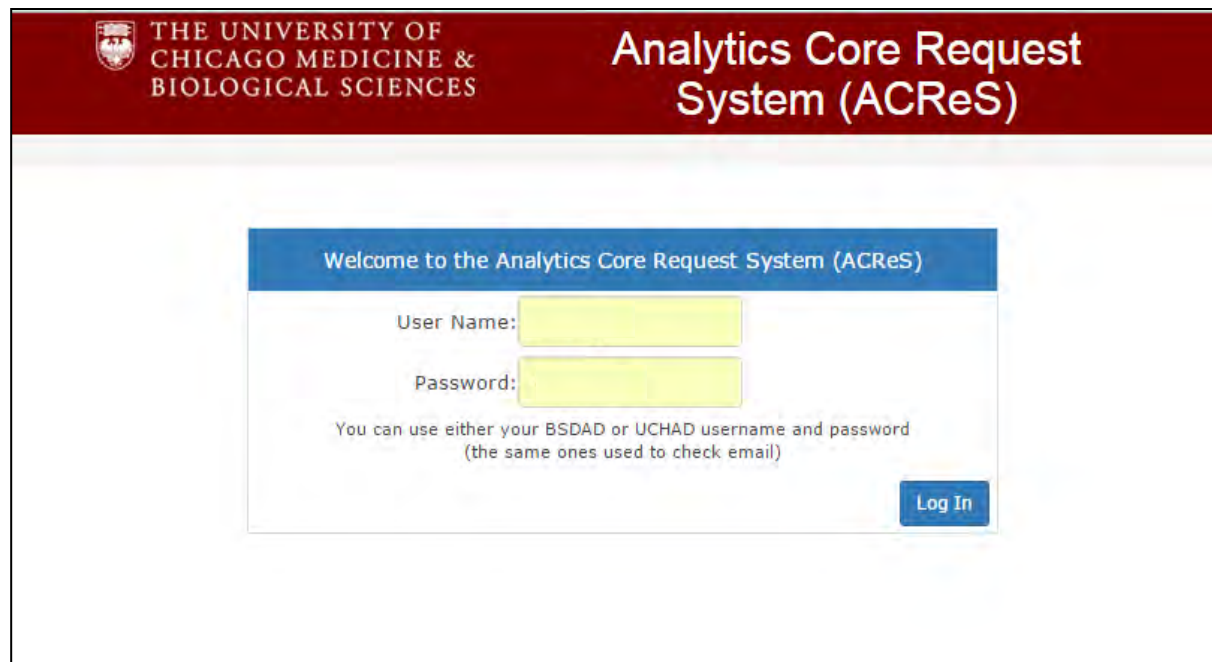# Current UCM Analytics Ecosystem

- Analytics Core

  - CBIS
    - Clarity
    - Epic reporting work bench
    - Epic Programming
  - Center for Quality(CFQ)
    - Advanced Analytics and Data Science
    - Clinical Effectiveness analytics
    - Health Care Delivery Science and Innovation
  - Center for Research Informatics (CRI)
  - Finance
    - Managed Care and Financial Planning
    - Budget Resource Analysis & Financial Planning
  - Patient Care Services
  - Procedural Services
  - UCPG decision support
  - Marketing Analytics
  - Strategic Planning

How Do I Request Data? ACReS

- Analytics Core Request System

- https://cri-app02.bsd.uchicago.edu/ACReS

Step 1: Log into ACReS Website

- https://cri-app02.bsd.uchicago.edu/ACReS
- Analytics Core connects the organization to data beyond the Center for Quality

Step 2: Begin a New Request

Analytics Core Request System (ACReS)

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

My ACReS Requests   New ACReS Request   Log out

↻ Copy Information from Previous ACReS Request

Requester name (First Last)

Samuel Volchenbou

Requester e-mail address

svolchen@peds.bsd.uchicago.edu

Requester Department

svolchen@peds.bsd.uchicago.edu

Description of problem, and purpose of request

This is a test request to demonstrate the system.

Request Urgency

Standard

Do you have an IRB # or are you applying for one?

Yes

Will this report contribute to knowledge usable outside of UCM?

No

Are there intentions for these data to leave the organization, outside of the University of Chicago Medicine?
(UCMC,University BAA templates)

No (if this changes please contact Analytics Core)

Step 3: Complete the Form

What is the refresh frequency of this report?

Never, just needed once ▼

Data types needed:
(select all that apply)
- ☐ Benchmarking (UHC, Compdata)
- ☑ Clinical (e.g. flow sheet, orders, meds, vitals)
- ☐ Employee / Staff
- ☐ Facility billing (e.g. MSDRG, ICD9 codes)
- ☐ Financial / Budget
- ☐ Patient Experience (e.g. survey responses)
- ☐ Patient Flow (e.g. arrival, transfer times)
- ☐ Physician billing (e.g. CPT codes)
- ☐ Other (enter response) / I don't know

Final output preference:
(select all that apply)
- ☑ Table (patient list, monthly rates)
- ☐ Grouped by (geography, patient, provider, time)
- ☐ Statistics (average, percent, SD, regression, simulation)
- ☐ Visualization (bar, pie, line chart, request guidance, unique)
- ☐ Control chart (e.g. p-, c-, or individuals chart)
- ☐ Other (enter response) / I don't know

Patient safety
(3) Known issue;
(2) suspected;
(1) low impact
- ○ 3 ◉ 2 ○ 1 ○ N/A

Compliance
This report is needed for:
(3) Regulatory agency mandate;
(2) UCMC standard;
(1) 'nice to have'
- ○ 3 ○ 2 ◉ 1 ○ N/A

Step 4: Submit the Form

Efficiency/ Productivity
Opportunity to patient throughput, labor, and/or non-labor is:
  (3) Known and central to this request;
  (2) suspected;
  (1) low impact
    ○ 3 ◉ 2 ○ 1 ○ N/A

Scholarly Activity
Research or educational dependency on this report is:
  (3) High;
  (2) moderate;
  (1) low dependency
    ○ 3 ◉ 2 ○ 1 ○ N/A

Financial
Total potential impact is estimated at:
  (3) >$500k;
  (2) >$100k;
  (1) <$100k
    ○ 3 ○ 2 ○ 1 ◉ N/A

Strategic goals
This is a goal at the level of:
  (3) Organization, AOP;
  (2) Chief/VP;
  (1) Executive Admin, Director
    ○ 3 ○ 2 ○ 1 ◉ N/A

Have you already identified or been in discussion with a lead analyst or member of the Analytics Core?

  No ⇕

Will this report require Protected Health Information (PHI)?

  Yes ⇕

[Submit] [Clear] [Cancel]

Step 5: Post Request Submission

- You will receive a confirmation email summary

- Check back to see status updates (you will receive updates via email, and through the ACReS tool)

- Your request will be assigned to the most qualified analytics team by the end of the next business day

- The analytics team assigned to your request will now be your primary contact for this request

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

Regulatory Compliance for Data Requests

QI vs. Research

Human Subject Research

- Human subjects research is defined in CFR §46.102 as
  1. Research: a systematic investigation … designed to develop or contribute to generalizable knowledge.
  2. Human subject: living individual about whom an investigator conducting research obtains
     a. Data through intervention or interaction with the individual, or
     b. Identifiable private information.

Quality Improvement

- Quality improvement is sometimes defined in the following ways:
  1. "Quality improvement (QI) consists of systematic and continuous actions that lead to measurable improvement in health care services and the health status of targeted patient groups."1
  2. Quality improvement is "the combined and unceasing efforts of everyone - healthcare professionals, patients and their families, researchers, payers, planners and educators - to make the changes that will lead to better patient outcomes (health), better system performance (care) and better professional development"2
  3. The pursuit of the triple aim: "Improving the U.S. health care system requires simultaneous pursuit of three aims: improving the experience of care, improving the health of populations, and reducing per capita costs of health care." 3

REFERENCES: (to outside sources, if applicable)

1. U.S Department of Health and Human Services, Health Resources and Services Administration. Quality improvement. 2011. Accessed 29 October 2015 at http://www.hrsa.gov/quality/toolbox/508pdfs/qualityimprovement.pdf.
2. Batalden P, Davidoff F. What is ''quality improvement'' and how can it transform healthcare? *Qual Saf Health Care* 2007;16:2–3
3. Berwick D, Nolan TW, Whittington J. The Triple Aim: care, health, and cost. *Health Affairs,* 27, no.3 (2008):759-769.

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

Regulatory Compliance for Data Requests

QI vs. Research

- **Quality Improvement Determination**
  - *New policy* that guides the approval process for the request of and use of institutional data in quality improvement projects when there is intent to share the data outside the OHCA (Organized Health Care Arrangement)
  - OHCA consists of UCM, BSD, and Pritzker
  - Request and/or use of PHI is subject to minimum necessary requirement and may be subject to approval from Patient Compliance Office *(applies to all non-research requests for data related to treatment, payment, operations, and QI)*
- **IRB Review**
  - Governs human subjects research
  - IRB approved protocol required prior to beginning data request
  - Request and/or use of PHI follows minimum necessary requirement and is subject to IRB approval

Regulatory Compliance for Data Requests

QI vs. Research

- **Quality Improvement Determination**
  - Are not assigned to the CRDW
- **IRB Approval**
  - Are assigned to the CRDW

Neither QI Determination nor IRB approval is a guarantee of analyst time. The determination simply designates which analytics team is assigned to review your project for formal acceptance.
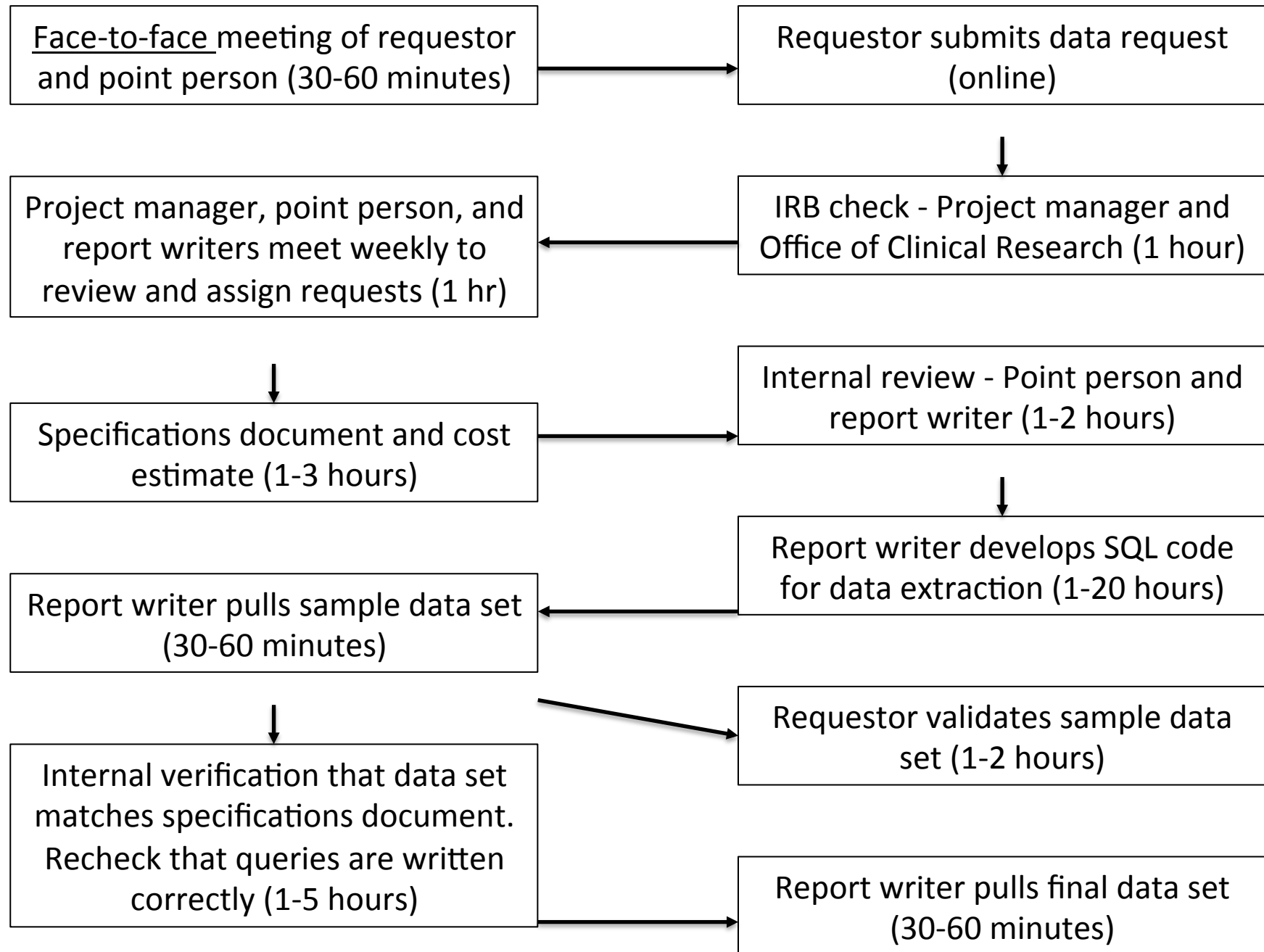
CRDW Request Triage

```
┌─────────────────────────────┐        ┌─────────────────────────────┐
│ Face-to-face meeting of      │───────▶│ Requestor submits data       │
│ requestor and point person   │        │ request (online)             │
│ (30-60 minutes)              │        │                              │
└─────────────────────────────┘        └─────────────────────────────┘
                                                      │
                                                      ▼
┌─────────────────────────────┐        ┌─────────────────────────────┐
│ Project manager, point       │◀───────│ IRB check - Project manager  │
│ person, and report writers   │        │ and Office of Clinical       │
│ meet weekly to review and    │        │ Research (1 hour)            │
│ assign requests (1 hr)       │        │                              │
└─────────────────────────────┘        └─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐        ┌─────────────────────────────┐
│ Specifications document and  │───────▶│ Internal review - Point      │
│ cost estimate (1-3 hours)    │        │ person and report writer     │
│                              │        │ (1-2 hours)                  │
└─────────────────────────────┘        └─────────────────────────────┘
                                                      │
                                                      ▼
┌─────────────────────────────┐        ┌─────────────────────────────┐
│ Report writer pulls sample   │◀───────│ Report writer develops SQL   │
│ data set (30-60 minutes)     │        │ code for data extraction     │
│                              │        │ (1-20 hours)                 │
└─────────────────────────────┘        └─────────────────────────────┘
              │                    ┌──▶ ┌─────────────────────────────┐
              ▼                    │    │ Requestor validates sample   │
┌─────────────────────────────┐   │    │ data set (1-2 hours)         │
│ Internal verification that   │───┘    │                              │
│ data set matches             │        └─────────────────────────────┘
│ specifications document.     │
│ Recheck that queries are     │        ┌─────────────────────────────┐
│ written correctly (1-5 hours)│───────▶│ Report writer pulls final    │
│                              │        │ data set (30-60 minutes)     │
└─────────────────────────────┘        └─────────────────────────────┘
```

CRDW Spec Document

PI:
Contact:
Title:
Request #: 11242
IRB#:
Data Type: PHI

Cohort Identification:
- Outpatients seen in Heme/Onc
- 4/1/2015-6/30/2015
- Designation of "palliative" intent in treatment plan

Index encounter is the first encounter that meets this criteria.

Data Set Output: Outpatient, Emergency, and Inpatient Encounters from Index through 6/30/2016

Cohort Info:
- MRN
- DOB
- DOD
- Race
- Gender
- Ethnicity
- Primary Language (as available)
- Religion (as available)

Encounter Info:
- MRN
- Bill_Num
- Enc_EIO
- Admission_type (planned, emergent, urgent)
- Adm_DTTM
- Dis_DTTM
- Start_DTTM
- Attending_Provider
- Visit_location
- Admit_location
- Dis_location
- Fin_class
- Marital_status
- Primary_ICD9_Dx
- Primary_ICD9_Desc
- Primary_ICD10_Dx
- Primary_ICD10_Desc
- Flag field for ICU admission during encounter

Diagnosis Info:
- MRN
- Bill_Num
- ICD9_Dx
- ICD10_Dx
- Charlson Comorbidity Score

Procedure Info:
- MRN
- Bill_Num
- ICD9_Px
- ICD10_Px
- CPT

Medication Info:
Anti-neoplastic agents only
- MRN
- Bill_num
- Med_name
- Order_DTTM
- Start_DTTM
- End_DTTM
- Given_DTTM
- DC_DTTM

Radiation Therapy:
- MRN
- Bill_num
- DTTM (of occurrence)

Notes: Index Encounters
ECOG score via Natural Language Processing
- MRN
- Bill_Num
- ECOG score
  (Could also be PS: 0-4)

Cancer Registry Info:
- MRN
- Date of initial diagnosis
- Primary site
- Date of first contact
- Current treatment

Time Estimate (Standard; 4-6 weeks):

Task Description	# Hrs	
Initial meeting at Office Hours	1	Not billed
Generate, document, and review data request specifications	1	Not billed
Generate SQL queries to pull requested data	22	
Cancer Registry	6	
Analysis: NLP	10	
Create, QA, and deliver sample/final data set	2	
Billable Hours ($95/hr)	40	$ 3800.00

Data Request Challenges: Most Common Hold-ups

- *Regulatory*
  - Distinction between data points used for screening vs. those needed for the actual data set, particularly related to PHI

- *Lack of response*
  - Unwillingness to meet during office hours
  - Timeliness responding to data clarification emails

- *Cohort identification*
  - Data is not available in a format that is conducive to data extraction

- *Request expectations*
  - Asking us to making the binary decisions
  - One patient per row

- *Funding*
  - No budget to facilitate the research

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

CRI Resources

- **Office Hours**
  - Tuesdays, Room N161, by appointment 10-4
  - Fridays, Room N161, by appointment 10-12

- **IRB guidance**
  - CRDW specific IRB language

- **ITM grant application guidance**
  - Website

- **Semi-self service tools**
  - I2b2
  - SEE Cohorts

- **Data Storage Solutions**
  - REDCap
  - Bulkstorage

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

CRI Resources

Semi Self-Service Tools

- i2b2

Semi Self-Service Tools

- SEE-Cohorts: https://seecohorts.cri.uchicago.edu/