THE UNIVERSITY OF CHICAGO MEDICINE

The Clinical Research Data Warehouse

Julie Johnson, MPH, RN

Senior Healthcare Business Analyst, CRI

Timothy Holper, MA, MS

Director of Data Warehousing and Business Intelligence, CRI

May 3, 2018

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 available at UCM but are not meant to provide an exhaustive analysis of all data sources

- It is not permissible to share these materials outside UCM

Agenda

- What is a Data Warehouse?
- What is the CRDW?
- CRDW I/O
- Data Warehousing Challenges
- Data Request Challenges
- Regulatory Compliance
- Making a Data Request
- CRI Resources
- SEECohorts
- Questions?

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
  - EMR: billing codes, orders, procedures, flowhseets, notes, labs, meds, etc.

- **Integrated**: Removal of inconsistencies regarding naming convention and value representations
  - E.g. HGB, hemoglobin, Hemoglobin all indicate the same laboratory test and the value for that variable is numerical with xx number of characters.

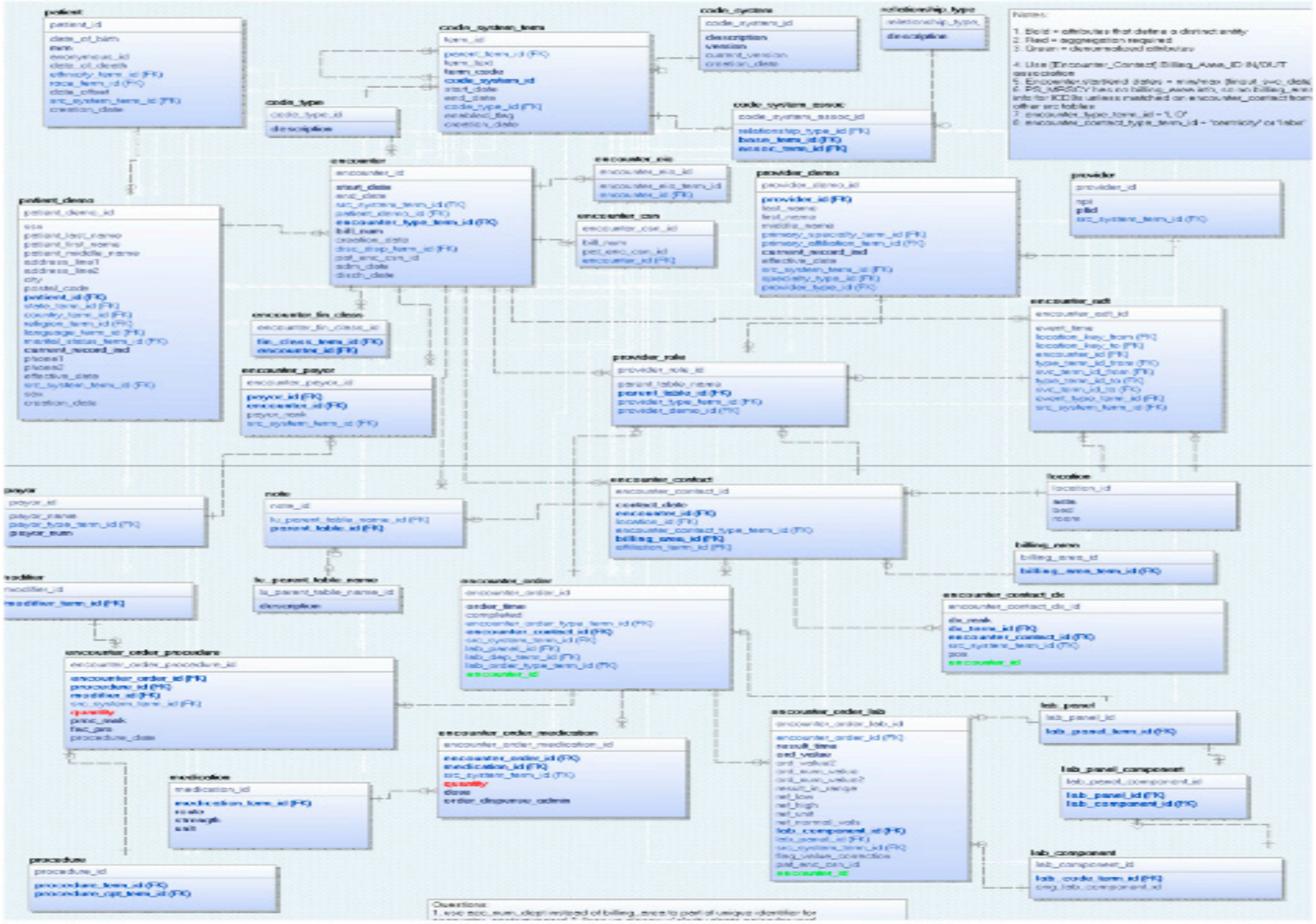
- **Nonvolatile**: Data stored in read-only format and updated daily

- **Time Variant:** Data not in real-time
  - Back-up of the original data; data may change over time as values are updated.

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 research questions

Data Warehouse is a Physical Database

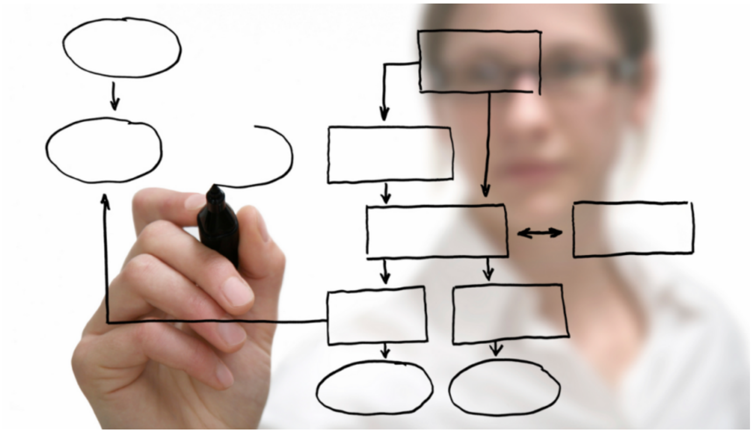


Data warehousing is a process

More than just hardware and software

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

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**



Data Warehousing is a process…

- Which requires a team…

What is the CRDW?

What is the CRDW?

- CRDW = Clinical Research Data Warehouse

- CRDW is a repository of University of Chicago medical data dating back to 2006.

- Process that brings together data from disparate sources, including Epic electronic medical records, Centricity billing, Cancer Registry, and REDCap, to create cohesive datasets for research.

- CRDW is an interface for ‘seamlessly’ integrating clinical and billing data with other data sources such as REDCap, Cancer Registry, and LabVantage

What is IN 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)

CRDW: By the Numbers

Subject Area	Record Count
Patients	991,780
Providers	83,222
Encounters	12,628,239
Diagnoses	28,378,974
Medications	54,110,714
Procedures (CPT)	56,733,136
Labs	406,945,626

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, usually just cost-prohibitive

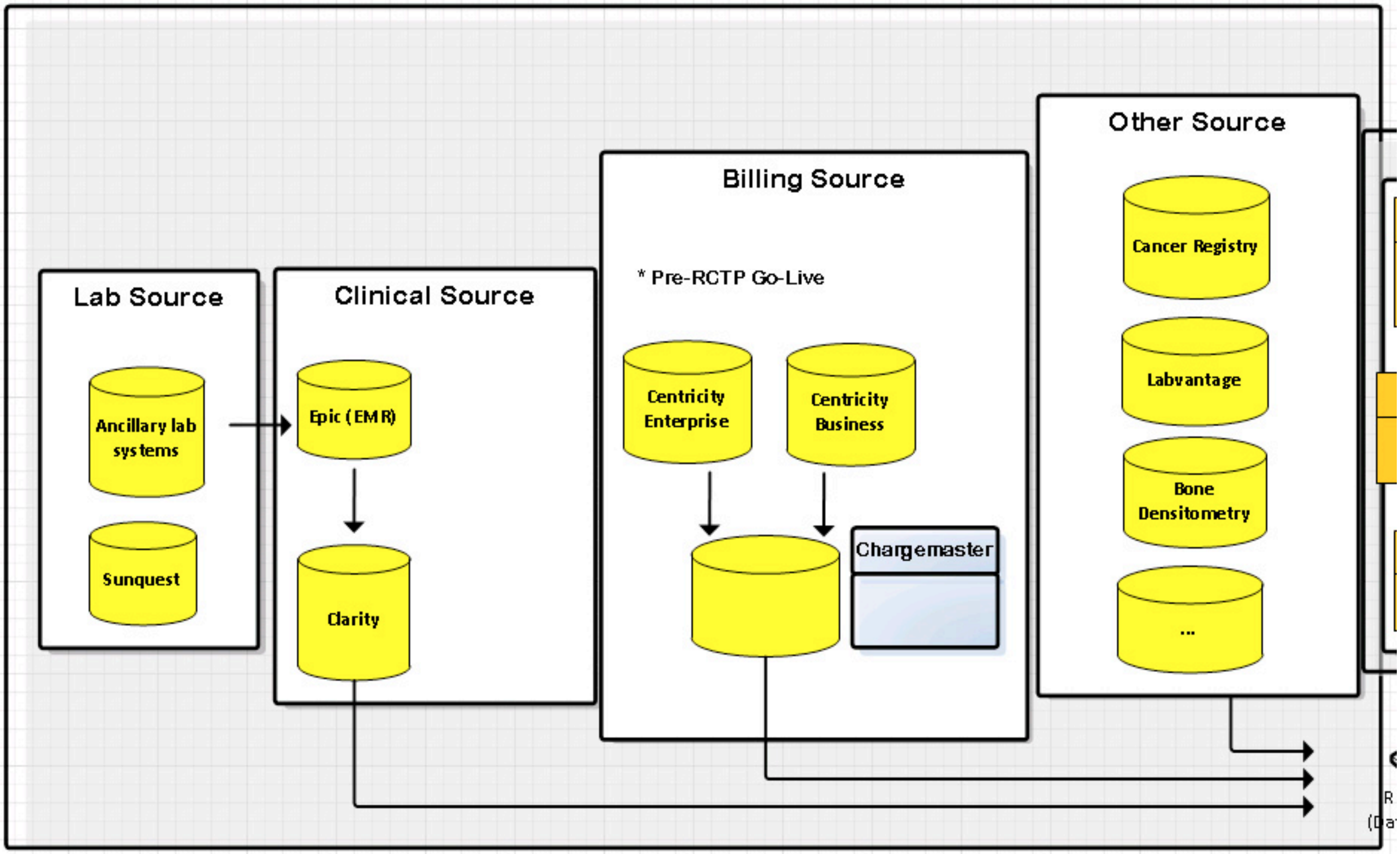
CRDW I/O

Source Systems- **Internal** to UCM

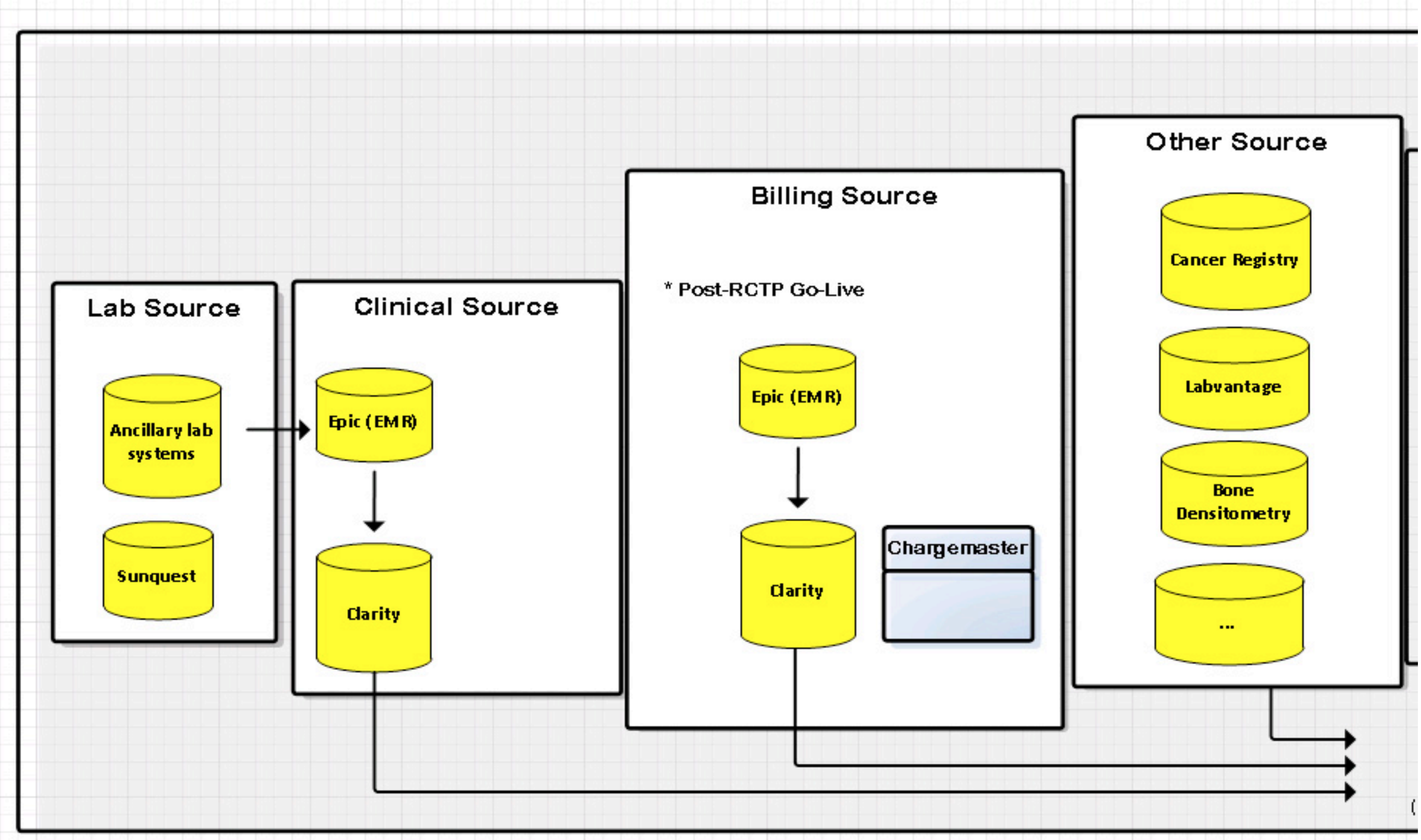
- UCM source data lives within a number of disparate source systems around the University of Chicago Medical Center

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

Source Systems- **Internal** to UCM



Source Systems- **Internal** to UCM



Source Systems- **External** to UCM

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

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

- 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.

De-Identification

- Definition found in section CFR 164.514 (b)(2)(i) of HIPAA. A de-identified data set is one in which either:
  1. The 18 identifiers specified in section 164.514(b)(2)(i) have been removed and the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify the individual (safe harbor method);
  2. A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, determines the risk is very small that the information could be used by the recipient, alone or in combination with other reasonably available information, to identify an individual (section 164.514(b)(1)), and documents the basis for such determination.
- De-identified Data will be used and disclosed in compliance with UCMC Policy A05-22 and BSD Policy PC66.

De-Identification

Using ‘Safe Harbor’ Method, we remove the following data elements from a data set:

- Names
- All geographic subdivisions smaller than a state (i.e. street address, city, county, ZIP Code)
  - except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Facsimile numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web universal resource locators (URLs)
- Internet protocol (IP) address numbers
- Biometric identifiers, including fingerprints and voiceprints
- Full-face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

De-identification Example (Discrete)

mrn	pt_name	gender	race	ethnicity	date_of_birth	date_of_service	address	postal_code
1234567	Smith, John	M	White	Not Hispanic or Latino	4/5/1950	1/15/2006	123 Pine Street	60618

mrn	patient_id
1234567	1

patient_id	date_offset
1	-13

patient_id	pt_name	gender	race	ethnicity	date_of_birth (offset)	date_of_service (offset)	address	postal_code (redacted)
1		M	White	Not Hispanic or Latino	3/23/1950	1/02/2006		606

Data Warehousing Challenges

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 obtain information about this encounter, such as **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: Structured vs. Unstructured Data

- Researcher puts data into EMR 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 (i.e. Doc Flow sheets, Templated Notes)
  - 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 and requires NLP

Example: Researcher wants to define cohort based on extracting patient's ECOG score from note

- Although the ECOG score is a discrete value, it may not be stored as a discrete field
- Difficult to extract

Data Warehousing Challenges: De-identification

- Potential for missing an identifier

  1. Overlooking a field which IS clearly one of 18 PHI elements
  2. Overlooking a value buried within a field which IS NOT clearly one of 18 PHI elements
     - Example #1: 18 Lab components in Clarity w/ name = 'Order Date'
     - Example #2: 3 Flow Sheet Measures w/ name = 'Patient Name'
     - …
- Recommendation for better Reference Data Management (RDM) to help minimize this.
- As Honest Brokers, maintaining and protecting mappings
- …

Data Request 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
- Just b/c you can do something doesn't mean you should
- A few suggestions:
- 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).

Regulatory Compliance

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.

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 (and any other UC staff, student, or faculty that is conducting work related to health care)
  - 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

In addition to regulatory requirements, data must follow UCM/BSD security standards (i.e. encrypted computers, HIPAA compliant storage, etc.)

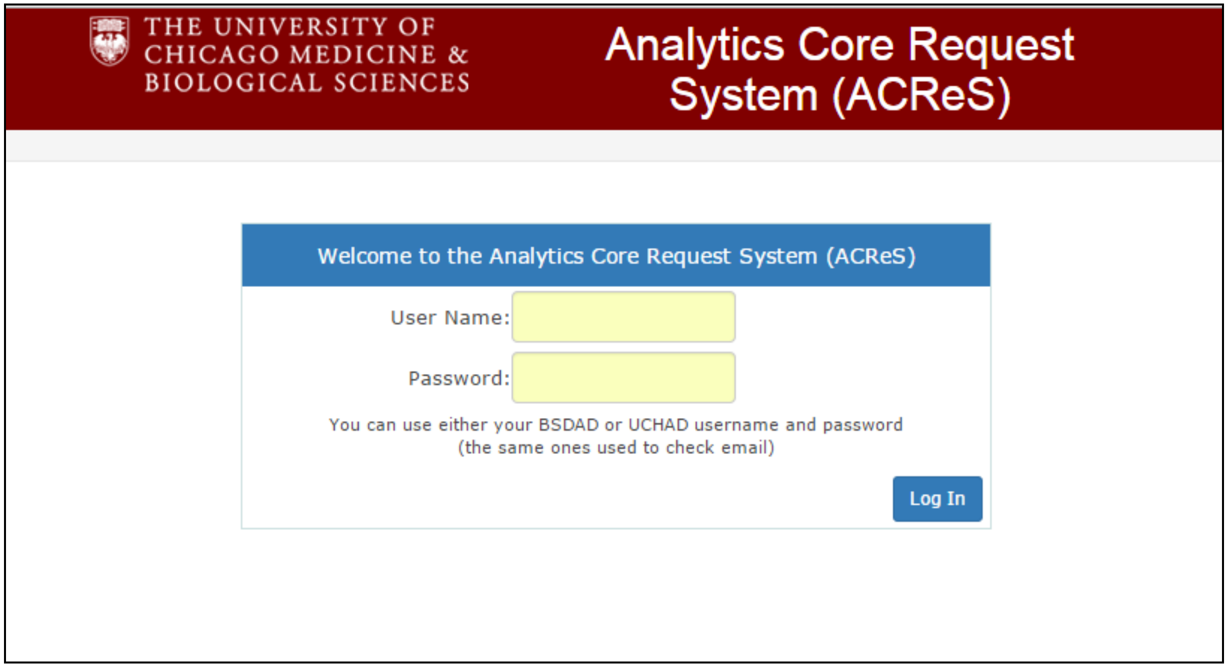
Making a Data Request

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



THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

Analytics Core Request System (ACReS)

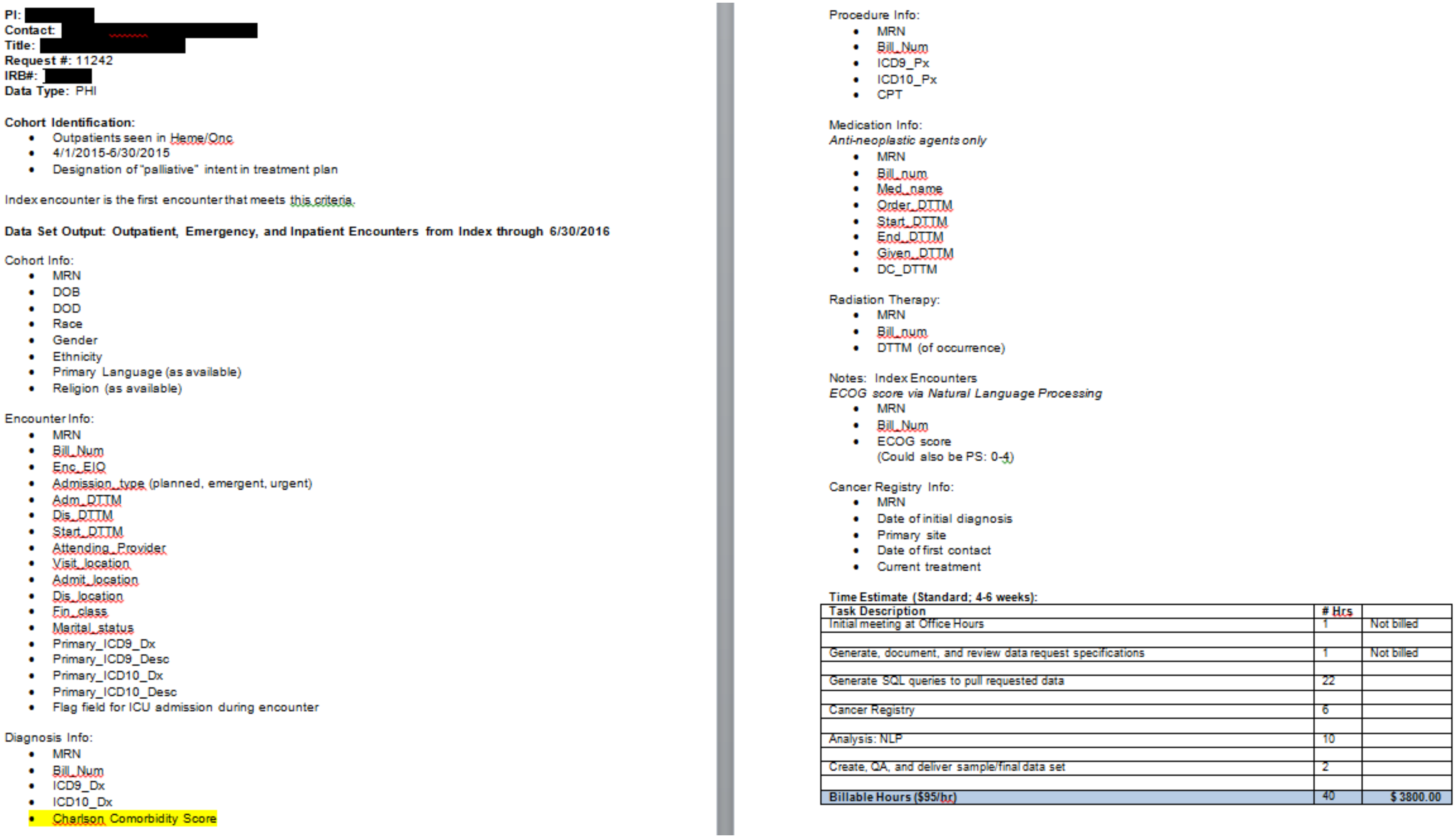
Welcome to the Analytics Core Request System (ACReS)

User Name:

Password:

You can use either your BSDAD or UCHAD username and password
(the same ones used to check email)

CRDW Spec Document



PI: [redacted]
Contact: [redacted]
Title: [redacted]
Request #: 11242
IRB#: [redacted]
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**

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 (AKA Affiliated Data Mart): regularly updated data
  - 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

CRI Resources

CRI Resources

- **Office Hours**
  - Tuesdays, Room N161, by appointment 10-4
  - Fridays, Room N161, by appointment 10-12
  - Other days as available

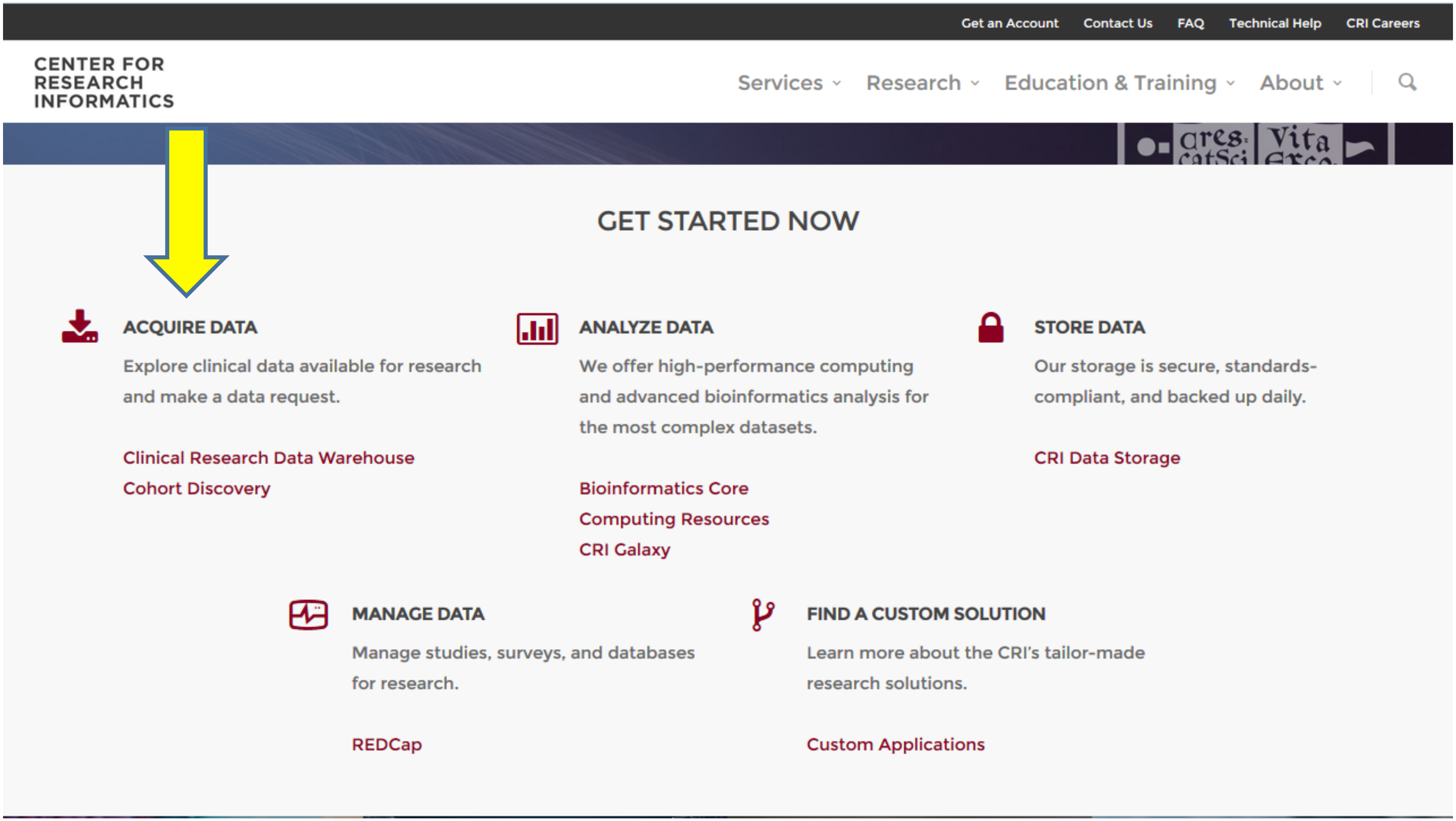
- **IRB guidance**
  - CRDW specific IRB language

- **Grant application guidance**

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

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

CRI Resources



GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources
CRI Galaxy

STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

CRI Data Storage

MANAGE DATA

Manage studies, surveys, and databases for research.

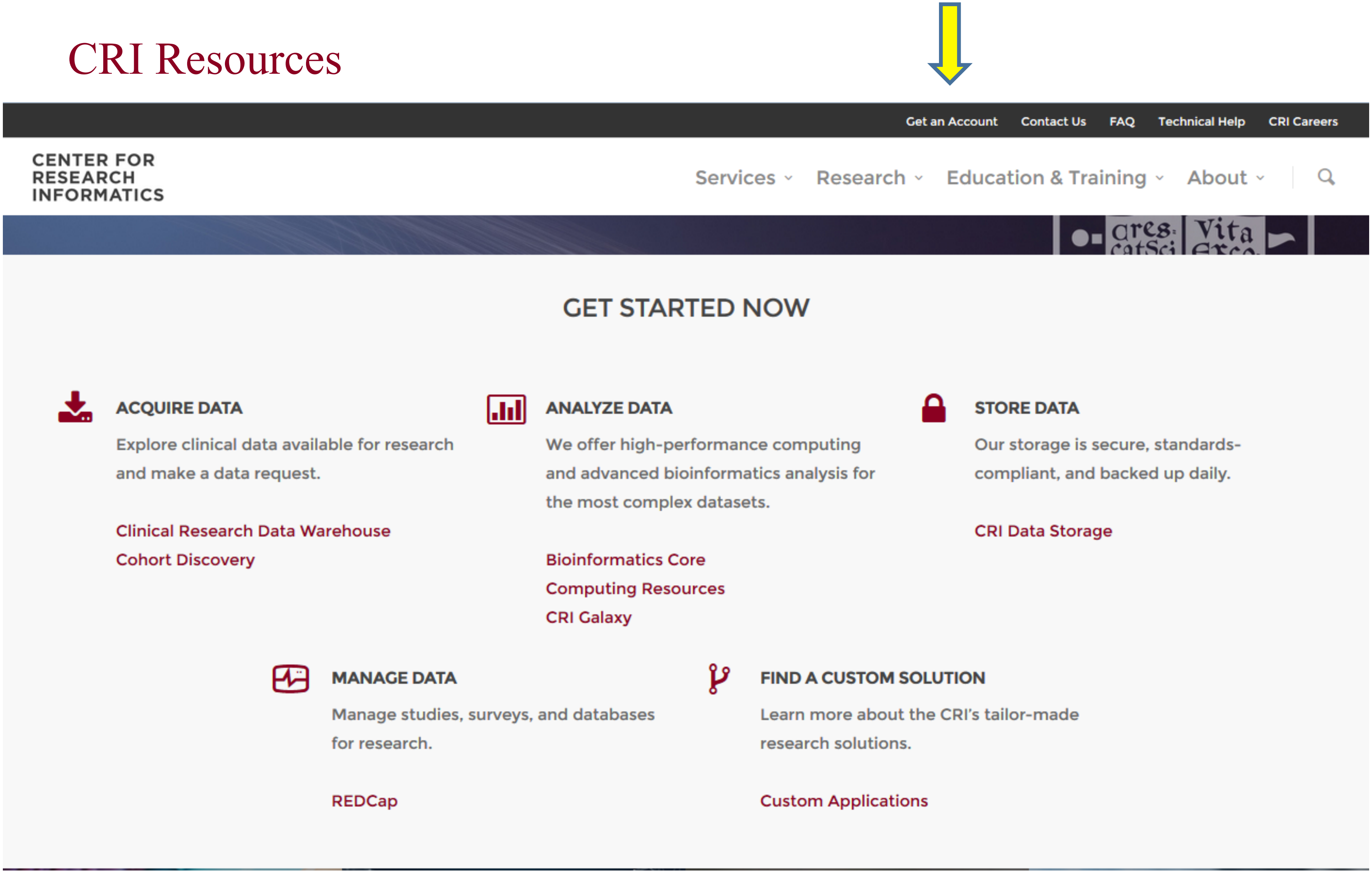
REDCap

FIND A CUSTOM SOLUTION

Learn more about the CRI's tailor-made research solutions.

Custom Applications

CRI Resources



GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources
CRI Galaxy

STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

CRI Data Storage

MANAGE DATA

Manage studies, surveys, and databases for research.

REDCap

FIND A CUSTOM SOLUTION

Learn more about the CRI's tailor-made research solutions.

Custom Applications

SEE Cohorts

SEECohorts Access

SEECohorts User Access

- Please turn on your computer and navigate to: https://seecohorts.cri.uchicago.edu
- Log in using your BSDAD or UCHAD credentials as noted on the right hand side of the screen
- If you do not have an account, your login information is incorrect, or you cannot log in despite having correct credentials, please let me.

- BSDAD\kdemanelis
- BSDAD\sbesser
- BSDAD\lkozloff
- UCHAD\lkozloff
- BSDAD\clyttle
- BSDAD\smaron
- UCHAD\smaron
- BSDAD\falenghat
- UCHAD\falenghat
- BSDAD\aflores
- BSDAD\bpaterso
- BSDAD\aorourke
- UCHAD\aorourke
- BSDAD\jluke1
- BSDAD\yyu
- UCHAD\mming

SeeCohorts:
Assessing Feasibility and Exploring your Cohort

- Navigate to https://seecohorts.cri.uchicago.edu and log in.

- Sample Query Parameters
  - Black, men, aged 25-45, hypertension by ICD9/ICD10
  - White, female, aged 18-89, Friedreich's Ataxia by ICD9/ICD10

  - Steps:
    - Drag and drop cohort search parameters into box on right
      - Boxes are joined by 'And' at same encounter
      - Within boxes = 'Or' at same encounter
    - Select 'Execute Query'
    - Encounter Parameters
      - Rename query
      - Select cohort criteria to view
    - Select Build the cohort
    - Explore a patient by selecting an encounter to view

Extra Slides

De-identification (Notes)

- De-identification according to the safe harbor method using software that locates suspected PHI elements within free text and replaces these identifiers with specific redaction tags.
- Isolating suspected PHI within text (including part of speech detection, dictionary lookup, pattern matching, etc.) is consistent with industry best-practices and current peer-reviewed research in text de-identification

De-identification Example (Notes)

GENERAL ADULT 48-HOUR DISCHARGE NOTE This form is only to be used for problems of a minor nature and for patients who require a 48 hour or less period of hospitalization. Name **NAME[AAA BBB]

MRN **ID-NUM

Sex Male

Date of Birth **DATE[Jul 16 1989]

Age 23Yrs

Attending Physician **NAME[ZZZ] Resident / **NAME[VVV] / PA **NAME[XXX]

Admission Date **DATE[Apr 13 2013]

Discharge Date/Time **DATE[Apr 14 2013] 1:28 PM REASON FOR ADMISSION Asthma exacerbation HOSPITAL COURSE HPI 23 yo M w/ asthma presents complaining of SOB, chest tightness since this AM. Pt reports increased coughing for the last few days, with increasing SOB and requirement for controller medicines last night. He had difficulty sleeping due to SOB. When he woke up this AM, he was acutely SOB and immediately came to ED. Pt has been using taking Symbicort, Singulair daily, but ran out of Symbicort 1wk ago. His symptoms have not responded to prn albuterol today. His baseline peak flow is >400, this AM in ED peak flow ~200. He denies f/c, n/v, abd pain, confusion, headache, lower extremity swelling. He does endorse some cough and rhinorrhea for the past couple days. Has been intubated >9x previously for severe asthma exacerbations, most recently in 2012. Last asthma exacerbation 2 months ago. He states that he comes to ED ~20x/yr for asthma. Brief Course Pt started on steroid burst with PO prednisone and continuous nebs in the ED. Symptoms improved with this therapy. He was then tapered to q6 duonebs overnight. The following day he no longer had any wheezing on exam. TREATMENT RENDERED AsthmaExacerbation: improved - continue duonebs q6 - prednisone 40mg daily burst x5 days - restarted home symbicort and singulair - provided Rx for refill of rescue albuterol inhaler ADVERSE REACTIONS None CONDITION ON DISCHARGE Pt was no longer SOB at time of discharge. He was provided with prescriptions for symbicort and his rescue albuterol inhaler. He states that he has enough of his other medications at home. He was scheduled to follow up with me in clinic in a few weeks. I will consider referral to pulmonology for management given his severe asthmahistory. **NAME[WWW CCC XXX], M.D. Internal Medicine, PGY-1 Pager ***PHONE

De-identification Challenges (Notes)

- Volume of data
  - Documents in corpus plus tokens in documents
- Note formatting is often valuable in determining whether a token is PHI, but formatting is stripped in Clarity ETL
- Ambiguity of natural language
  - “Foley cathether” vs. “John Foley, MD”
- Variability in date coding affects date shifting
  - “1/23/2017” vs “Jan 23, 2017” vs “January 2017”
- No silver bullet
  - There is an inherent risk vs utility tradeoff in selecting the eagerness of the algorithm

Distributed Data Models: CAPriCORN Project

- CAPriCORN is a clinical research data sharing initiative involving 11 Chicago-area institutions
- Common data model devised and each institution is responsible for instantiating a local version of the data model, writing ETL code to populate this instance with data from their local data warehouses, and making this available to the network for querying cohort counts
- A major hurdle to interoperability of data between sites in the network is the frequent lack of mapping from local code sets to shared ontologies for labs and medications
- Patient de-duplication across sites is accomplished by means of a hashing algorithm which takes as input various combinations of demographic fields and produces a hashed identifier; given identical input fields, the algorithm will produce the same hashed identifier which will allow for tracking a single patient across multiple sites

Salient Points

- How data goes into system dictates how well we can pull data out (i.e. garbage in, garbage out)
- Technology will not, by itself, fix bad data or bad business processes
- Must start w/ standardized data models, widely accepted business rules, 'clean' data, and governance processes when pushing data into a system and pulling data out of a system
- Good data governance and data warehousing allows us to collaborate with other institutions and share data