Weill Cornell Medicine NewYork-Presbyterian

High Fidelity:

Connecting information for Better Research Reproducibility

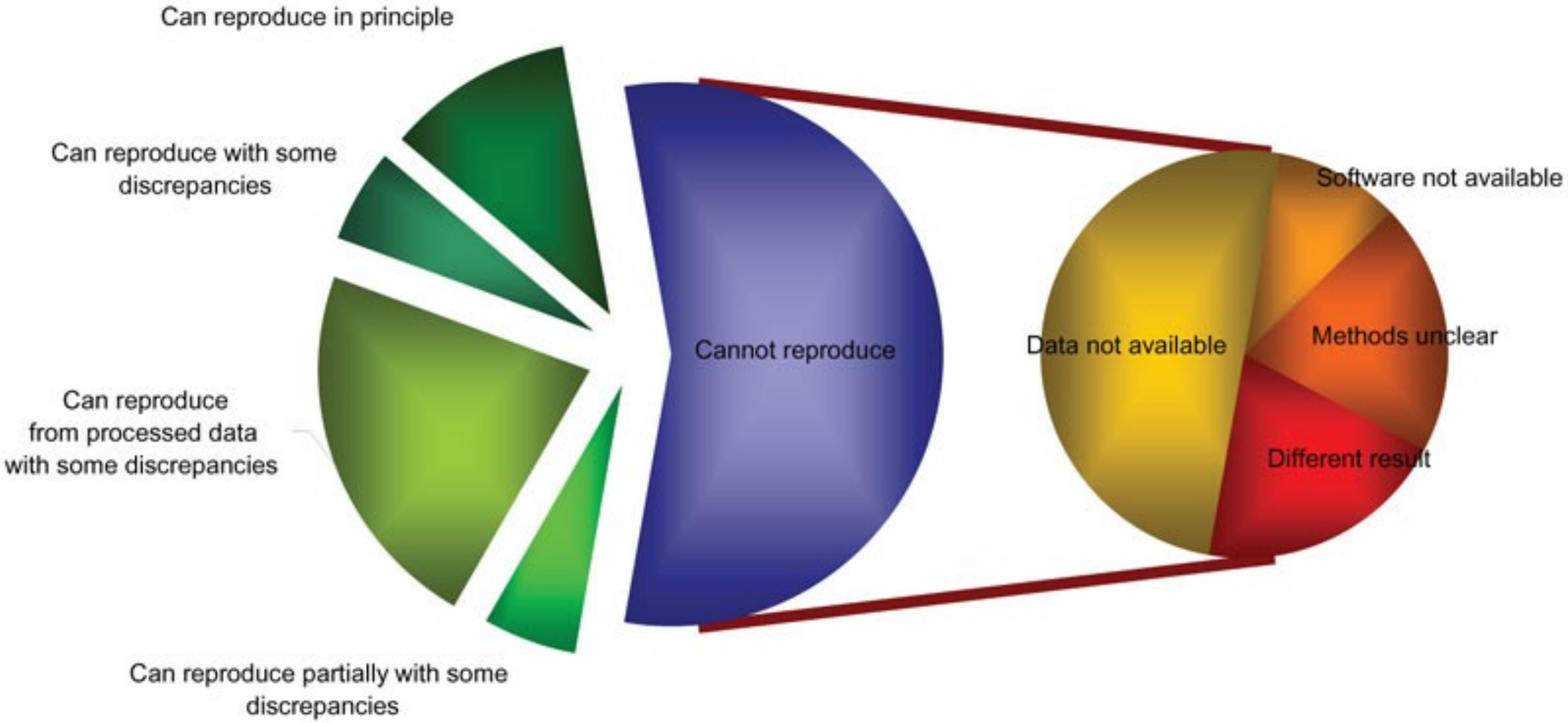
Terrie Wheeler, AMLS
Director
Samuel J. Wood Library

Peter Oxley, PhD
Associate Director for Research Services
Samuel J. Wood Library

There are three fires burning that motivate our actions towards better reproducibility

1. “Reproducibility Crisis”
2. Data Retention Mandates
3. Allegations of Misconduct

Attempting to reproduce research proves to be difficult, when you can’t get the data…



Ioannidis *et al.* (2009) Nat. Gen. 41:149

“OMB Circular A-110 states that the retention period is th
from the date the final financial report is submitted.”

“NSF states in its General Grant Conditions that records must be retained
for three years after the submission of all required reports”

“in the case of research misconduct involving NIH fund
records must be retained for six years
after the final resolution date of the case.”

“retain research data pertinent to patented inventions
for the life of the patent”

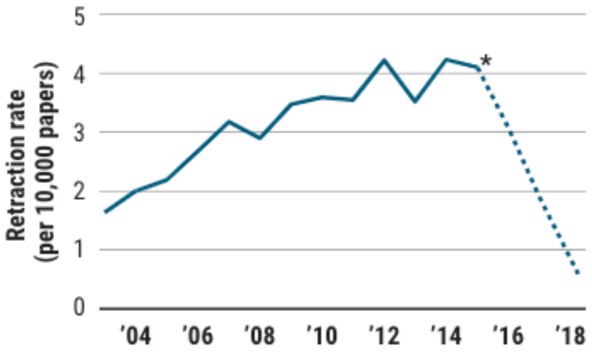
“about 2% of scientists admitted to have fabricated, falsified or mod
data or results at least once”

“Up to one third admitted a variety of other questionable practices”

Fanelli (2009) PLOS One 0005738

“3.8% of published papers contained problematic figures, with at lea
exhibiting features suggestive of deliberate manipulation”

Bik *et al.* (2016) mBio 00809-16



Brainard and You (2018) Science 00809-16

There are three fires burning that motivate our actions towards better reproducibility

1. "Reproducibility Crisis"
2. Data Retention Mandates
3. Allegations of Misconduct

Maximizing data value and ethical research conduct

We are not automatically making research reproducible, but providing the infrastructure that helps make research re

Reporting Standards

- PRISMA
- MIAME
- MIQE

Principles of reproducibility

- FAIR
- Literate Programming (Knuth 1992)
- Compendiums (Gentleman and Lang, 2003)

Data Publication standards

- Open Science

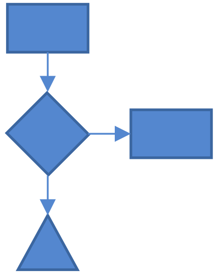
Communities of practice

- The Turing Way

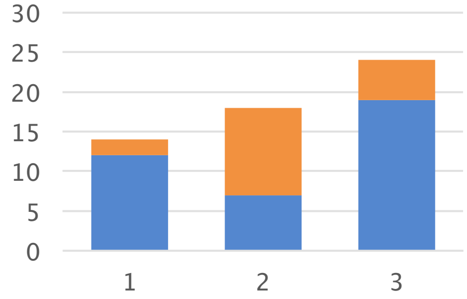
How can we help researchers capture the data and workflows that lead to publishable results?

010101000101
010111100001
001101001001
000111111001
010001111010
001010100100
100100101111

DATA



WORKFLOWS



RESULTS

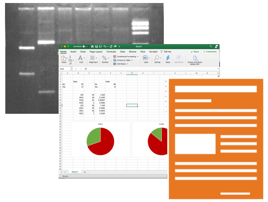
Hurdle #1: Dealing with confidential data

Hurdle #2: Researchers are very busy

Hurdle #3: Data and workflows are very diverse

Hurdle #4: Maintaining a high quality solution

Piece #1: Electronic Lab Notebooks capture (small) data and while remaining flexible to researcher needs



Direct storage of images, data files, analysis files, workflows

File versioning and immutable timestamps

Integration with Jupyter Notebooks (on roadmap)

Shareable and transferable

Piece #2: A file management system for tagging, tracking, a then archiving files on the institutional storage systems

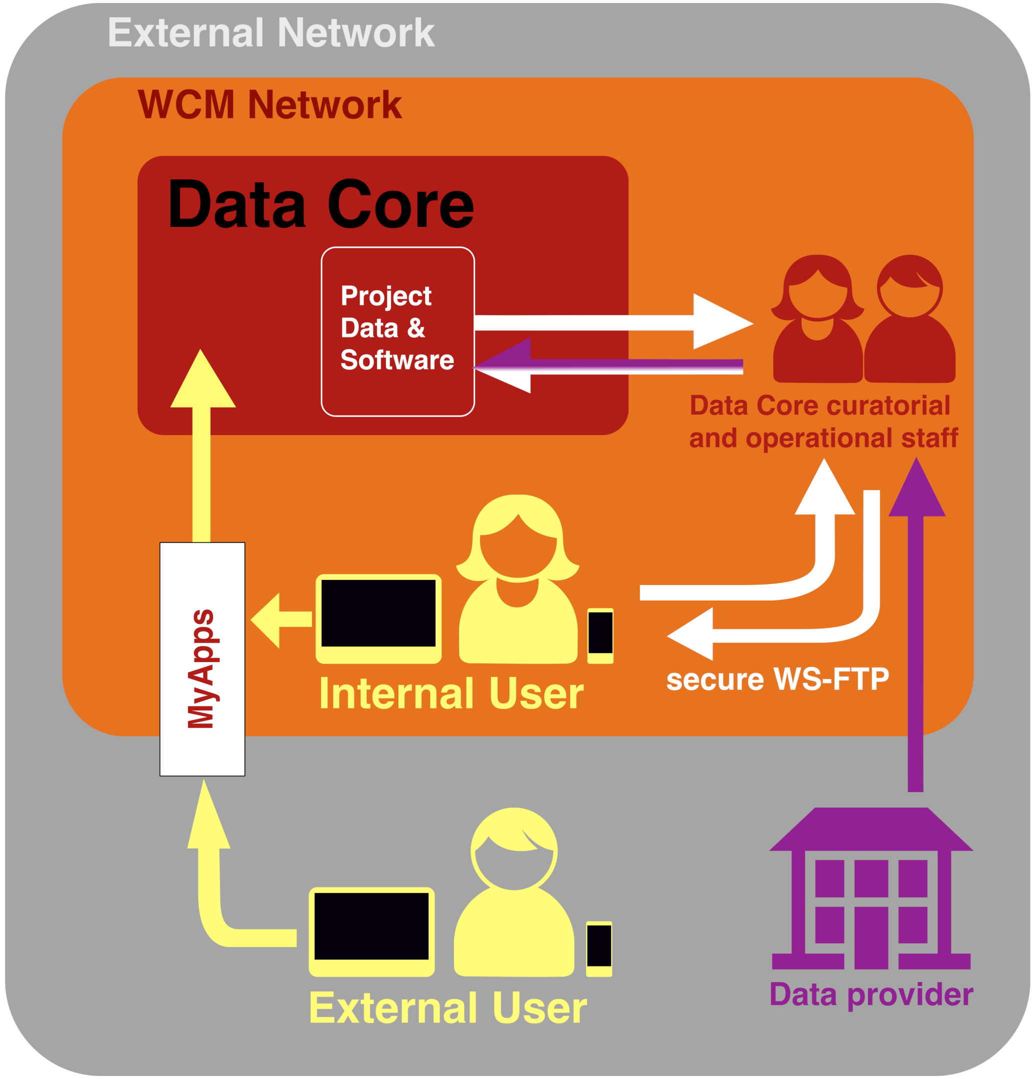
Data associated with a project/publication assigned a unique ta

File identity managed by hashing

File location tracked within institutional storage

Actionable scripts for marking project complete, and archiving

Piece #3: Secure file access management and computation t the institutional Data Core

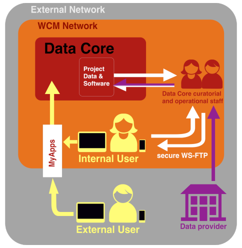


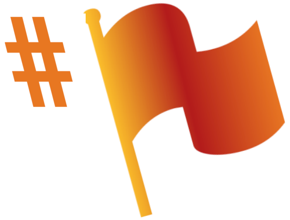
Secured, collaborative, flexible

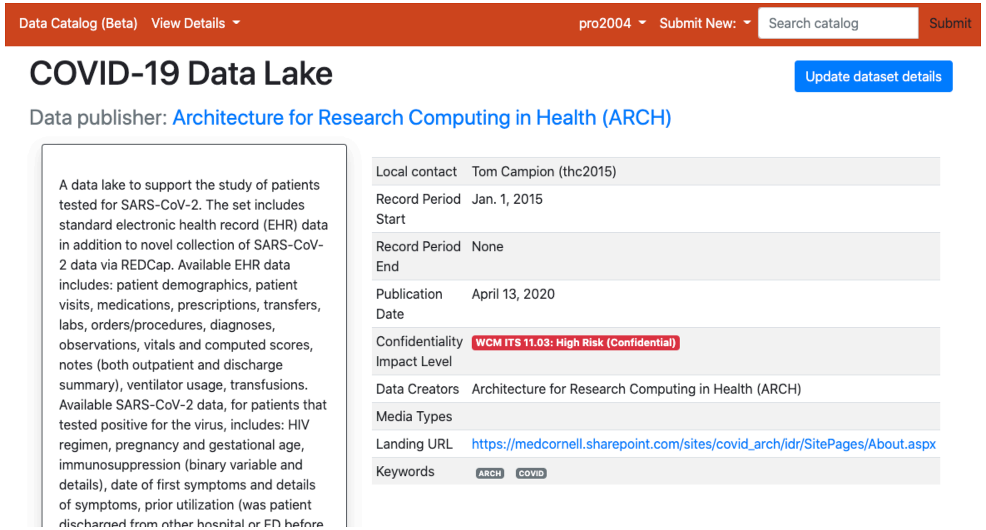
Project governance and monitoring

Data curation for import and export

Holding the pieces together: an institutional Data Catalog that connects data, workflows, governance and access condi







COVID-19 Data Lake

Data publisher: Architecture for Research Computing in Health (ARCH)

A data lake to support the study of patients tested for SARS-CoV-2. The set includes standard electronic health record (EHR) data in addition to novel collection of SARS-CoV-2 data via REDCap. Available EHR data includes: patient demographics, patient visits, medications, prescriptions, transfers, labs, orders/procedures, diagnoses, observations, vitals and computed scores, notes (both outpatient and discharge summary), ventilator usage, transfusions. Available SARS-CoV-2 data, for patients that tested positive for the virus, includes: HIV regimen, pregnancy and gestational age, immunosuppression (binary variable and details), date of first symptoms and details of symptoms, prior utilization (was patient discharged from other hospital or ED before

Local contact	Tom Campion (thc2015)
Record Period Start	Jan. 1, 2015
Record Period End	None
Publication Date	April 13, 2020
Confidentiality Impact Level	WCM ITS 11.03: High Risk (Confidential)
Data Creators	Architecture for Research Computing in Health (ARCH)
Media Types	
Landing URL	https://medcornell.sharepoint.com/sites/covid_arch/idr/SitePages/About.aspx
Keywords	ARCH COVID

Discovery layer with the capacity to connect grants, data, publicatio

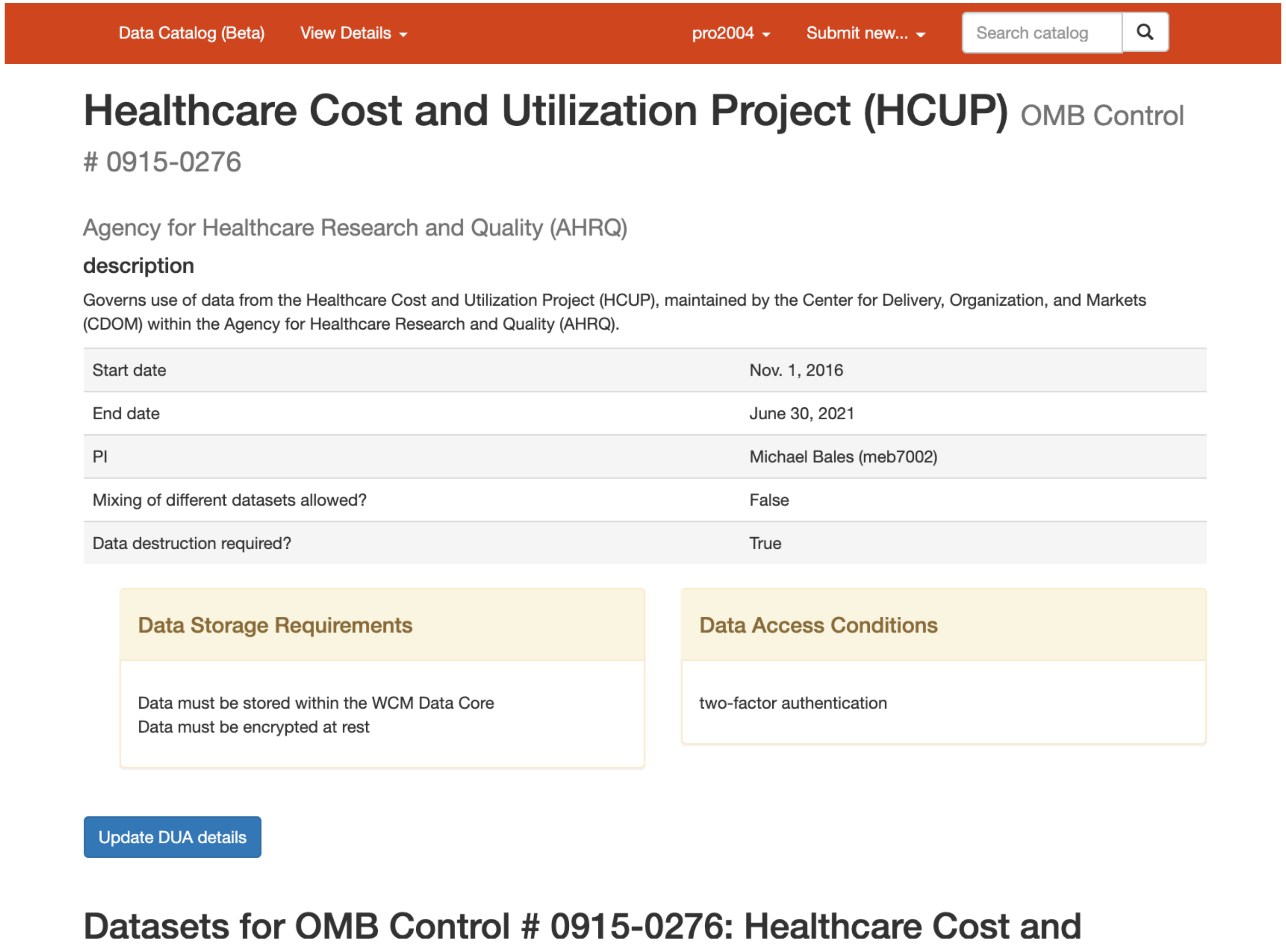
The WCM Data Catalog was built to manage data governance and access conditions

Scope of authorization

User authorization

Data Controls

Reuse scope



Healthcare Cost and Utilization Project (HCUP) OMB Control # 0915-0276

Agency for Healthcare Research and Quality (AHRQ)

description

Governs use of data from the Healthcare Cost and Utilization Project (HCUP), maintained by the Center for Delivery, Organization, and Markets (CDOM) within the Agency for Healthcare Research and Quality (AHRQ).

Start date	Nov. 1, 2016
End date	June 30, 2021
PI	Michael Bales (meb7002)
Mixing of different datasets allowed?	False
Data destruction required?	True

Data Storage Requirements

Data must be stored within the WCM Data Core
Data must be encrypted at rest

Data Access Conditions

two-factor authentication

Datasets for OMB Control # 0915-0276: Healthcare Cost and

Three triggers to prompt capture and storage

1. Project/Grant completion
2. Publication
3. Faculty member leaves the institution

Move files to archive
Register project, file tag, ELNs in Data Catalog

Acknowledgements:

Curt Cole (WCM CIO)

Data Core

Alice Chin
Ana Proper
Frank Ashmun
John Ruffing
Jo Hargitai
Lucy Walle
Michael Bales
Heather Kleinschmidt

Electronic Lab Notebooks

Marie Linvill (LabArchives)
Cindy Chen
Danny Tan
Lucy Walle
Tony DiFazio

Data Catalog

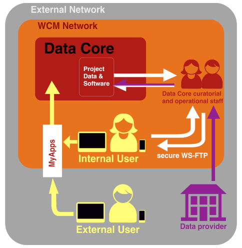
The Data Catalog Collaboration (DDC)
Nicole Contaxis (NYU Langone)
Tom Campion

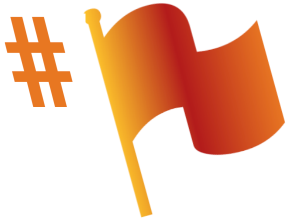
File System Management

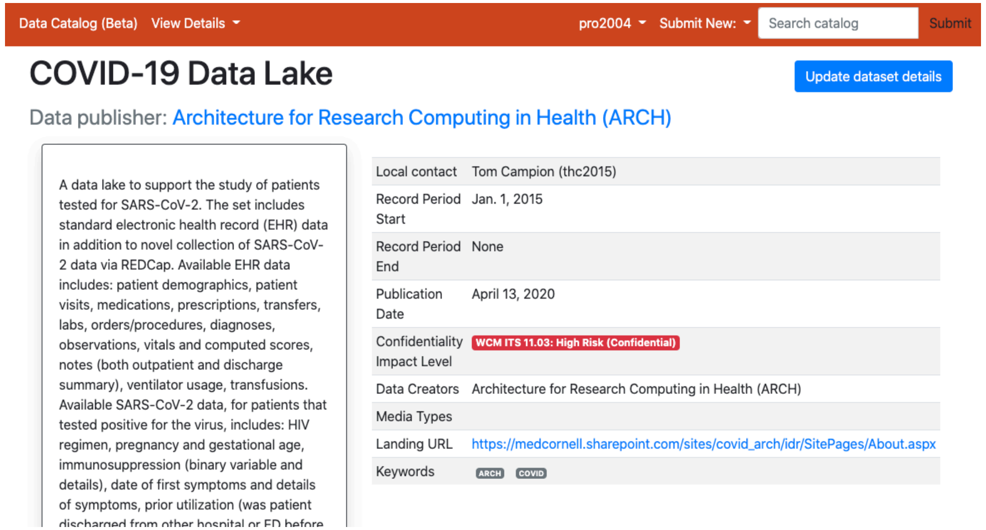
Cindy Chen
Darrin Stivala (Starfish)
Doug Hughes (Starfish)
Jo Hargitai
Tony DiFazio

Peter Oxley pro2004@med.cornell.edu
Terrie Wheeler tew2004@med.cornell.edu

Design and Implementation of a Secure Computing Environment for Analysis of Sensitive Data at an Academic Medical Center.
Oxley PR, Ruffing J, Campion TR Jr, Wheeler TR, Cole CL.
AMIA Annu Symp Proc. 2018 Dec 5; 2018:857-866.







COVID-19 Data Lake

Data publisher: Architecture for Research Computing in Health (ARCH)

A data lake to support the study of patients tested for SARS-CoV-2. The set includes standard electronic health record (EHR) data in addition to novel collection of SARS-CoV-2 data via REDCap. Available EHR data includes: patient demographics, patient visits, medications, prescriptions, transfers, labs, orders/procedures, diagnoses, observations, vitals and computed scores, notes (both outpatient and discharge summary), ventilator usage, transfusions. Available SARS-CoV-2 data, for patients that tested positive for the virus, includes: HIV regimen, pregnancy and gestational age, immunosuppression (binary variable and details), date of first symptoms and details of symptoms, prior utilization (was patient discharged from other hospital or ED before

Local contact	Tom Campion (thc2015)
Record Period Start	Jan. 1, 2015
Record Period End	None
Publication Date	April 13, 2020
Confidentiality Impact Level	WCM ITS 11.03: High Risk (Confidential)
Data Creators	Architecture for Research Computing in Health (ARCH)
Media Types	
Landing URL	https://medcornell.sharepoint.com/sites/covid_arch/idr/SitePages/About.aspx
Keywords	ARCH COVID