



## Clinical Data Warehouse for Research

### BMC Clinical Data Warehouse (CDW) for Research CRRO Series Part 1: Getting Research-Ready Data for Studies

May 11, 2022

# CDW for Research – Two Part Series

## Part 1:

### Getting Research-Ready Data for Studies

May 11, 2022

- Describe the BMC CDW for Research, data available in the data warehouse, services offered by the CDW for Research, how to engage with the CDW for Research, and answers to other FAQs
- Discuss information and approvals required to effectively request data from the CDW for Research
- Share updates on the CDW for Research's current initiatives and strategic goals

## Part 2:

### Foundational Data Literacy

Wednesday, June 8, 2022, 12:00-1:00 PM

- Build a fundamental understanding of data and data structures to increase data-literate decisions.
- Describe how to optimally leverage data.
- Discuss exemplar health informatics projects and research case studies that meaningfully leveraged available data.

**REGISTER AT:**

[tinyurl.com/CDWpart2](https://tinyurl.com/CDWpart2)



# Questions to consider as we present...

---

What are common **data needs or thematic areas** in your studies?

---

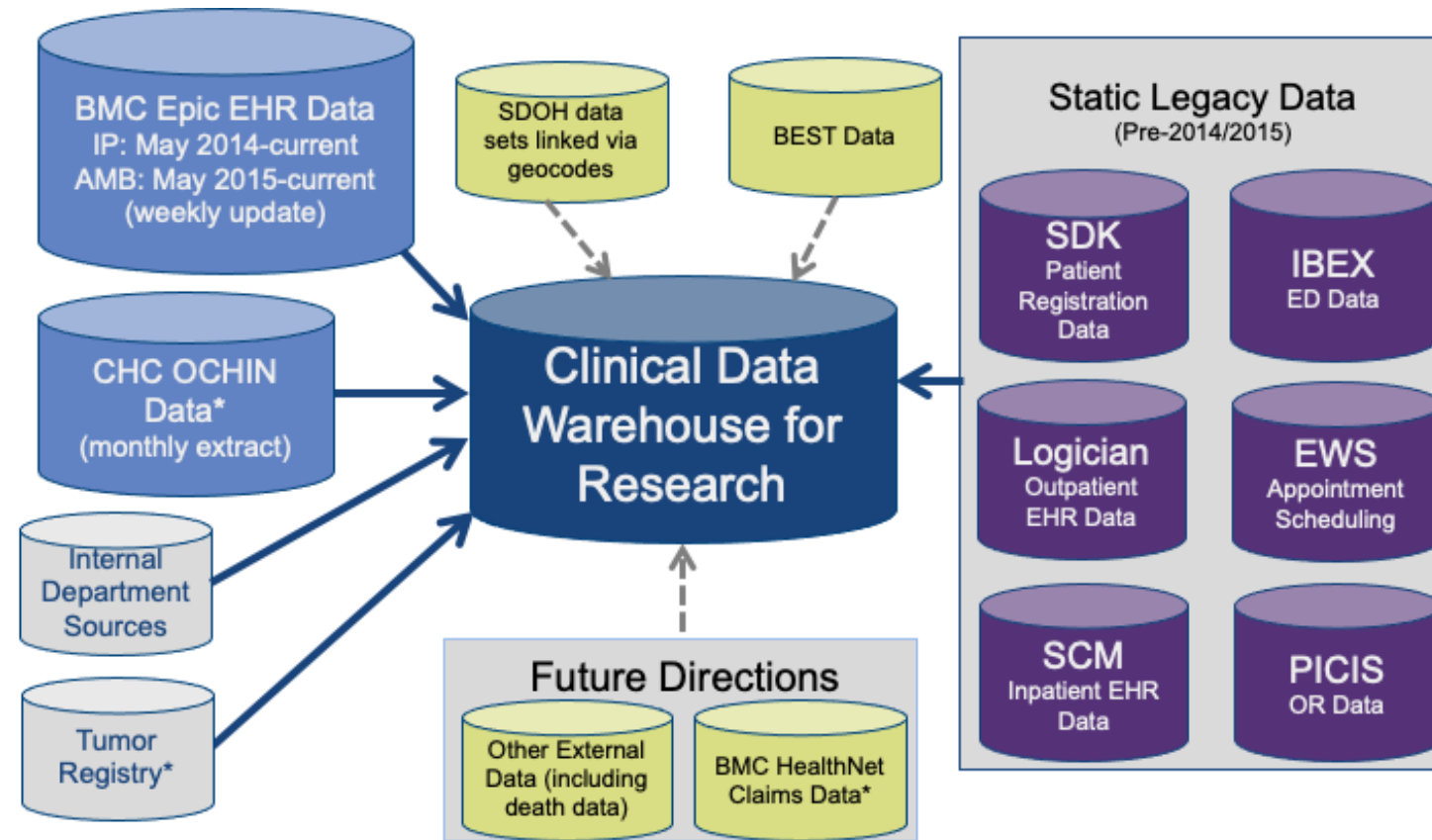
What are your **biggest struggles when getting data** for your studies?

---

What **resources or materials** would you like to see the CDW for Research create for research teams in the coming months?

# The BMC Clinical Data Warehouse for Research

- Centralized resource to **access patient-level and population-level data for research**
  - Electronic health records data (Epic)
  - Historical data from BMC's legacy clinical systems
  - CHC OCHIN EHR data
  - Piloting BMCHP claims data for research purposes
- Provides **simple aggregate counts** to prepare for grants or studies
- Provides **data extracts**
  - Cohort identification, recruitment lists, and data sets for prospective studies
  - Data sets for observational/retrospective studies
  - **Linked data sets** from multiple sources related to each other with **patient-level unique identifiers**



*\*special permission required to access*

*The CDW for Research also collaborates with Departments and Divisions to increase research infrastructure and better leverage data for research purposes.*

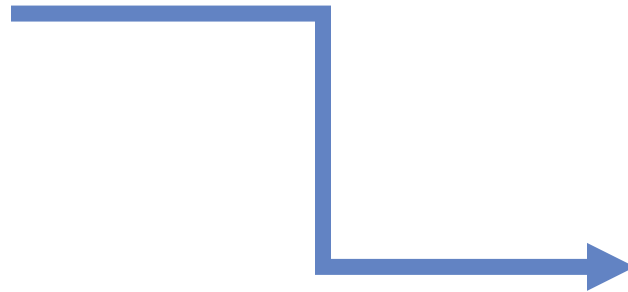
# Data Available in the Data Warehouse

- **Patient Information**
  - Demographics
  - Contact information
  - Social circumstances
- **Data collected during health care encounters**
  - Diagnoses
  - Lab values
  - Medications
  - Clinical notes
  - Data captured in smart forms or flowsheets
  - Encounter location/date/time
  - Data on clinical care team
- **Healthcare utilization**
  - Ambulatory visits
  - Telehealth visits
  - Emergency department utilization
  - In-patient admissions
- **Medical History**
- **Data from Services and Procedures**
  - Results
  - Surgical procedures
  - Image reports
  - Pathology reports
  - Radiology reports
- **Tumor Registry Information**
- **Administrative data**
  - Patient scheduling data
  - Epic billing data
  - Insurance information
- **And more...**

# Data Available in the Data Warehouse

General Rule

***Epic***



Clinical Data Warehouse

# Data Available in the Data Warehouse: SDOH & Health Equity

- **Race, Ethnicity, and Language (REaL) data**

- Primary Race
- Hispanic Indicator (Binary Y/N)
- Primary Ethnicity
- Primary Language
- Needs Interpreter Flag
- Previous Interpreter Usage

- **Gender, Sex, and Sexual Orientation data**

- Sex (assigned at birth)
- Gender Identity
- Sexual Orientation

- **Housing Insecurity**

- CDW for Research's housing insecurity algorithm
- BMC THRIVE screener – Housing question
- ICD-10 Z59.0 (Problems related to housing & economic circumstances)

- **Food Insecurity**

- CDW for Research's preventative food pantry referrals, letters, orders, and utilization algorithm
- BMC THRIVE screener – Hunger Vital Signs questions
- ICD-10s Z59.41 (Food insecurity) & Z59.48 (Other specified lack of adequate food)

- **Utilities Insecurity**

- BMC THRIVE screener – trouble paying for utilities
- BMC Utilities Shut-Off Protection Letter issuance

- **Country of Origin**

- Country of origin
- Place of birth
- Immigration status

- **Socioeconomic Indicators and Proxies**

- Employment Status
- Marital Status
- Education level / Years of Education
- Insurance as a Socioeconomic Proxy
- BMC THRIVE screener – Trouble paying for medications
- BMC THRIVE screener – Unemployed and looking for a job
- BMC THRIVE screener – Interested in more education

- **Geocoding**

- Geocoding/Census tract
- CDC Social Vulnerability Index (SVI)

# What people think data extraction is...

The Google logo is displayed in its standard multi-colored font.

BMC Patients with positive FIT test over past two years





# What data extraction is...

```
create table [redacted] nologging as
WITH
GET_DATA AS (
SELECT DISTINCT
  prov.prov_full_nm as fit_ordering_prov
  ,rslt.src_pat_id
  ,rslt.src_order_id
  ,rslt.RESULT_SPECIMN_TAK_DT
  ,rslt.result_dt
  ,dept.src_department_nm as department_nm
  ,ord.src_ord_auth_prov_id as prov_id
  ,rslt.result_value_txt
FROM [redacted] rslt
  INNER JOIN [redacted] prov on rslt.src_order_prov_id = prov.src_prov_id
  INNER JOIN [redacted] ord on rslt.src_order_id = ord.src_order_id
  INNER JOIN [redacted] dept on to_char(ord.SRC_ORD_DEPT_ID) = dept.src_department_id
  INNER JOIN [redacted] vis on ord.src_enc_id = vis.src_enc_id
WHERE
  rslt.stool_hemoccult_yn = 'Y'
  and
  rslt.result_value_txt in ('POS','NEG')
  and
  rslt.result_dt >= '2018-01-01 00:00:00'
  and
  (vis.vst_Class_nm is null or vis.enc_type_nm = 'Appointment' or vis.vst_class_nm in ('OUTPATIENT','SPECIMEN'))
)
SELECT
  GET_DATA.fit_ordering_prov
  ,GET_DATA.src_pat_id
  ,GET_DATA.src_order_id
  ,GET_DATA.RESULT_SPECIMN_TAK_DT
  ,GET_DATA.result_dt
  ,lag(GET_DATA.result_dt) over (partition by GET_DATA.src_pat_id order by GET_DATA.result_dt asc) as prev_result_dt
  ,lead(GET_DATA.result_dt) over (partition by GET_DATA.src_pat_id order by GET_DATA.result_dt asc) as next_result_dt
  ,GET_DATA.department_nm
  ,GET_DATA.prov_id
  ,GET_DATA.result_value_txt
  ,sum(case when result_value_txt = 'POS' then 1 else 0 end) over (partition by get_data.src_pat_id) as NUM_POS_FIT_TESTS
```



**Clinical Data Warehouse  
for Research**

## **CDW for Research Services**

# CDW for Research Services: Data Extracts

## Simple Counts

Provide aggregate counts for study planning, feasibility analysis, and grant/proposal submission.

## Data Extracts

Data extracted from the data warehouse for your study cohort, organized and provided in excel files.  
*(mitigating need for manual chart review)*

## Linked Data Extracts

Data extracted from the data warehouse linked to community health center (CHC) data, claims data, and other external data.

## Recruitment/Cohort Lists

MRN, Name, DOB, Contact Information, Demographics, Fields used in the cohort inclusion.  
*(as permitted by IRB protocol)*

## Department- or Division-wide Efforts

Describe specific patient populations, increase research infrastructure, and better leverage data for research.

# CDW for Research Services: What else we provide...

- Apply **study cohort inclusion and exclusion criteria** and ensure data set conforms with all research regulations
- **Assign study IDs**
- **Create and manage master codes**, with or without providing access to study team
- Provide **subsequent data extracts** for the same cohort (whether study team has a master code or not)
- Develop **recurring and automated data pulls**
- **Consult with study teams** to identify robust data to answer your research question, develop budgets, complete your study IRB protocol, and/or determine if manual chart review would be required for your study.





**Clinical Data Warehouse  
for Research**

# How to Effectively Request Data from the CDW for Research

# Checklist: What needs to be in-place before CDW for Research Can Extract Data?

## ✓ Identify all data fields needed to answer your research question

- Engage clinical colleagues who are familiar with clinical workflows and Epic.
- Reach out to [cdw@bmc.org](mailto:cdw@bmc.org) to schedule a consultation early in the research planning process.

## ✓ Secure IRB approval or determination

- Clearly describe what data you are using when, and for what purpose, in your IRB protocol.
- Pay close attention to the HIPAA section: List all data fields needed, list dates that include full data range (pre/baseline, study period, post/follow-up), and ensure appropriate access to PHI and master code as needed.

## ✓ Identify funding

- The CDW for Research operates as a BMC Research Core Facility.
- See our website for current hourly rate tiers.

## ✓ Submit a fully-complete CDW for Research Data Request Form

- Step-by-step instructions for completing a well-defined data request are available on our website.
- A clear and complete data request is critical to expediting data turnaround – Read all instructions!

## ✓ Confirmation of any Special Approvals required

**CDW for Research Website and  
Data Request Form**

[tinyurl.com/CDWwebsite](https://tinyurl.com/CDWwebsite)



# Data that Require Special Approvals to Access

## Community Health Center (CHC) Data\*

- Permission from each Boston HealthNet CHC is required to access CHC data
- Boston HealthNet has a centralized Project Request Form to request access: <https://www.bu.edu/ctsi/community-engagement/boston-healthnet-bhn/>

## BMC HealthNet Plan Claims Data

- Permission from the State (MassHealth) is legally required for all research on a project-by-project basis
- Requires BMC HealthNet Plan legal team involvement
- Email [cdw@bmc.org](mailto:cdw@bmc.org) with questions

## Vaccine Data

- Special permission is required if data are extracted from the Massachusetts Immunization Information System (MIIS), which contains data on vaccines administered outside of BMC
- Researchers may access data on vaccines given at BMC without special permission

*\*The CDW for Research has access to data from the following CHCs: Boston Health Care for the Homeless Program, Codman Square, DotHouse, Greater Roslindale, Mattapan Community, South Boston, South End, Upham's Corner.*

# Avoiding Common Data Request Pitfalls

## ✓ There is a **mismatch between the CDW for Research Data Request and the study IRB protocol**

- Common mismatches: Dates do not span full range of when data are needed; specific PHI not requested in IRB protocol (e.g., MRN or DOB); Part II (SUD clinic) data not reflected on study protocol.
- Study team may need a broader cohort from the CDW for Research than what is your final study population/analytic set.
- Consider timelines of diagnosis and treatment, e.g., Cohort is from Jan 2021 – December 2021, but patient was diagnosed with the disease 5 years before study period and received treatment in 2021
- Mismatches require an IRB amendment before we can start your data extraction.

## ✓ **Not including all required information** in CDW Data Request Form

- Provide all ICD-10s, CPTs, procedure names, report names, form names, labs, and medication names.
- Avoid acronyms.

## ✓ **Not specifying and defining all data fields** you want the CDW for Research to provide

- List all individual data fields in the request form.
- Define all data (e.g., 'Depression' by ICD-10 or PHQ score? Both?) and at what timeframe (age at the time of the emergency department visit).



# Avoiding Common Data Request Pitfalls

Review the CDW for  
Research website

Read all instructions  
on the CDW for  
Research Data  
Request Form

Email [cdw@bmc.org](mailto:cdw@bmc.org)  
for a consultation

**CDW for Research Website and  
Data Request Form**

[tinyurl.com/CDWwebsite](https://tinyurl.com/CDWwebsite)





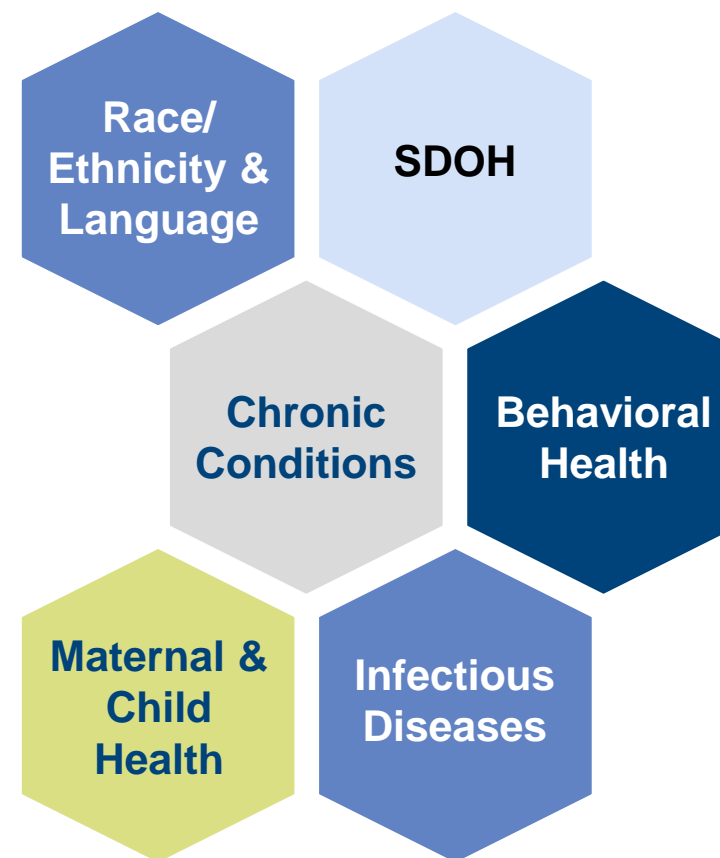
**Clinical Data Warehouse  
for Research**

# CDW for Research Initiatives and Strategic Goals

# Categorizing Key Patient Populations & Research Areas

- The CDW for Research develops **phenotypes for identifying patients, demographics, diseases, and characteristics** – improving consistency and data capture
- We assist investigators with **working with messy data** and making decisions about **how to accurately identify study populations**
  - Examples: How will you use EHR data to identify if a patient ...
    - ...is pregnant?
    - ...has a substance use disorder?
    - ...has housing insecurity?

## CDW for Research Focus Areas



# Patient Phenotyping: Obesity

- **Objective:** What proportion of the 2,729 patients diagnosed with COVID at BMC during spring 2020 were obese?
- **Approach #1:** Search CDW for all patients with ICD-10-CM codes for obesity or severe obesity in the active problem list, past medical history, or encounter diagnoses and a positive SARS-CoV-2 test → **N=350**
- **Approach #2:** Add measured BMI  $\geq 30$  to search criteria → **N=1,164**
- **Take home points:**
  - Adding measured BMI to search criteria yields >3x more patients with obesity
  - With approach #1, percentage with obesity = 13%; with approach #2, proportion with obesity = 43%
  - Implications for identification of obesity as a risk factor for COVID

Concretely defining patient populations and capturing all relevant data are essential to draw accurate conclusions.

# Patient Phenotyping: Experiencing Housing Insecurity

- **Objective:** What proportion of 2021 BMC patients\* experienced housing insecurity?
- **Approach #1:** Patients who screened positive for on either of the two BMC THRIVE screening tool housing questions → **N= 3,926 (1.9%)**
- **Approach #2:** Patients with ICD-10 Z59.0 (Problems related to housing & economic circumstances) in the medical record → **N= 5,423 (2.6%)**
- **Approach #3:** Patients identified as experiencing homelessness or housing insecurity using the CDW for Research housing insecurity algorithm (7 data points) → **N= 14,962 (7.1%)**
- **Take home points:**
  - There are often multiple ways to capture clinical phenotypes or conditions in the data, and study teams need to understand what data are 'most relevant' for their research question and study population to appropriately define variables.
  - CDW for Research has expertise in the data, and we can work with you to guide you in the definition process.

Understanding what you are trying to measure and why is crucial for identifying the 'best' data to answer your research question.

# BMC PATIENT SNAPSHOT – 2021

The data below are provided by the BMC Clinical Data Warehouse (CDW) for Research. Refer to the data dictionary on page 2 for data definitions and descriptions. Reach out to [cdw@bmc.org](mailto:cdw@bmc.org) for additional information.



Clinical Data Warehouse  
for Research

## 2021 Patient Population

This snapshot includes patients who had at least one completed office visit, telehealth visit, emergency department (ED) encounter, and/or inpatient encounter at BMC January 1, 2021 and December 31, 2021.

**Total # Patients with  $\geq 1$  encounter between 1/1/2021-12/31/2021 — 211,766**

### Primary Language (n=211,775)

English	147,635	70%
Haitian Creole	10,118	5%
Spanish	32,025	15%
Another Language	20,907	10%
Unknown	1,090	<1%

### Sex Assigned at Birth (n=211,775)

Female	117,119	55%
Male	94,592	45%
Unknown	64	<1%

### Race/Ethnicity (n=211,769)

Asian	8,913	4%
Hispanic or Latino	51,518	24%
Non-Hispanic Black or African American	69,132	33%
Non-Hispanic White	54,356	26%
Another Race or Ethnicity	2,429	1%
Declined or Unknown	25,421	12%

### Age (years)<sup>1</sup> (n=205,269)

0-1	3,737	2%
2-19	29,875	14%
20-44	80,287	38%
45-64	58,185	27%
65-79	26,156	12%
80+	7,029	3%

The CDW for Research uses SDOH-relevant data from Epic EHR and other data systems to support health equity research.



## GEOCODING & CENSUS TRACT DATA

- Patient-level geocodes, including census tract level
- Data can be linked to various social indices, including CDC Social Vulnerability Index (SVI)

## HOUSING INSECURITY ALGORITHM

- Seven flags, including ICD-10 codes, THRIVE screen, address history
- Creating a housing insecurity score



## FOOD INSECURITY ALGORITHMS

- Characterize patients that screen positive for food insecurity
- Accurately capture food pantry utilization

## OTHER LINKED SDOH DATA

- THRIVE data, Epic social history
- SUD phenotypes
- OUD & behavioral health curation
- Recently linked with Boston Emergency Services Team (BEST) data





Clinical workflows and clinical context have important implications when developing a data pull and interpreting extracted data.

Without clinically informed & intentional data practices, we risk inappropriate data use and inaccurate interpretation.

- Example: **Food Insecurity & BMC Preventative Food Pantry Utilization**
  - Food pantry referrals and utilization may be documented on behalf of a household.
  - Food pantry clinical workflow does indicate ICD-10s for food insecurity in the medical records; therefore, patients can be referred to or access the food pantry without an ICD-10 for food insecurity in their medical record.
  - Important to understand data implications to develop a data pull consistent with study research goals and questions.
- Example: **ED Utilization During COVID Pandemic**
  - Extracted data for a study based on anticipated clinical workflow.
  - Upon data analysis, the team sensed the utilization was misrepresented and realized the COVID+ patients were triaged differently and thus the mechanism to derive the cohort needed to be altered.
- Example: **Ectopic Pregnancies**
  - Variables were complex and required a bit of ‘digging’ to find the variables in the data warehouse.
  - The study team walked data analysts through the clinical workflow and clinical progression to understand the data capture and facilitates the analyst finding the ‘right’ data in the data warehouse.




# CDW for Research: FY22 Activities at a Glance

- Developing **resources for investigators and study teams** to aid in requesting and using data
  - Data dictionaries
  - SDOH data 'menu of options' guide
- **Collaborating with departments and divisions** to define specific patient populations and help research groups better leverage data.
- In process of **hiring 1.0 FTE analyst** to help cut-down on wait times for data from the CDW for Research.

# For more information CDW for Research services and fees, or to request data...

- ⇒ Visit our website: <https://www.bmc.org/research/clinical-data-warehouse-cdw>
- ⇒ Email: [cdw@bmc.org](mailto:cdw@bmc.org)



## BMC Clinical Data Warehouse (CDW) for Research

**Access Clinical Data for Research**

The Boston Medical Center (BMC) Clinical Data Warehouse (CDW) for Research leverages clinical information from BMC's Electronic Health Record (EHR) and other Health System-related data streams to create a repository of data to support research.

The CDW can:

1. [Help formulate](#) a CDW data request
2. [Estimate the cost](#) of a CDW data request
3. Provide count data for [feasibility analysis](#) or proposal/grant preparation without IRB approval (so called prep-to-research).
4. [Provide extracts](#) of data housed in the CDW for retrospective or prospective research studies. This includes cohort identification, individual-level data sets, and linked data sets from multiple sources. Projects requesting individual-level data must have IRB approval. Some data streams may require additional permission to access those data from the CDW, including OCHIN data from Community Health Centers.

**Current Wait Times & Fees**

Request Type	Current Wait Time	Fee
Simple Counts	Counts delivered within <b>10 business days</b>	No fee for aggregate, simple count requests
Cohort Identification and Recruitment List	List delivered in <b>2-4 weeks</b>	If funded by a BMC/BU account or grant, or if funds originate from a Boston HealthNet Community Health Center: <b>\$75 per hour</b> .
Patient-Level Data Sets	<b>6-8 weeks</b> before a CDW analyst can <b>begin</b> to work on a request. Research teams are encouraged to plan accordingly.	If funded by for-profit funders, or if funding is external to BU/BMC: <b>\$131.25 per hour</b>

**BMC Clinical Data Warehouse (CDW)**

Data Available from the CDW

Services

How to Request Data

Student, Resident, & Fellow Research

Fees and Billing

FAQs

Other BUMC Resources to Support Researchers

## Clinical Data Warehouse (CDW) for Research: Data Request Form

Use this form to submit a CDW research data request for research.

We encourage research teams to [visit our website](#) before submitting a CDW data request. Our website provides information on CDW services, data request form submission, and fees and billing.

### Current Wait Times & Fees

Request Type	Current Wait Time	Fee
Simple Counts	Counts delivered within <b>10 business days</b>	No fee for aggregate, simple count requests
Cohort Identification and Recruitment List	List delivered in <b>2-4 weeks</b>	If funded by a BMC/BU account or grant, or if funds originate from a Boston HealthNet Community Health Center: <b>\$75 per hour</b> .
Patient-Level Data Sets	<b>6-8 weeks</b> before a CDW analyst can <b>begin</b> to work on a request. Research teams are encouraged to plan accordingly.	If funded by for-profit funders, or if funding is external to BU/BMC: <b>\$131.25 per hour</b>

**Research teams must identify a source of funding before a data request will be added to the CDW work queue.**

If you are interested in speaking with CDW personnel about your study or request, email [cdw@bmc.org](mailto:cdw@bmc.org). **An initial consultation is provided to all studies free of charge.** Research teams do not need to submit a data request form to request a consultation.

*If you are concerned about CDW fees, please reach out to the CDW to schedule a consultation. We can work with you to keep costs low.*

Next Page

- ⇒ Submit a CDW for Research Data Request using our online form

---

What are common **data needs or thematic areas** in your studies?

---

What are your **biggest struggles when getting data** for your studies?

---

What **resources or materials** would you like to see the CDW for Research create for research teams in the coming months?