VA Research
DISCOVERY★ INNOVATION★ ADVANCEMENT
VA Informatics and Computing Infrastructure

Managing a Research Project Using VINCI Data

Olga Efimova and Andrew Holbrook
VA Informatics and Computing Infrastructure (VINCI) Resource Center
December 13, 2023

ChooseVA

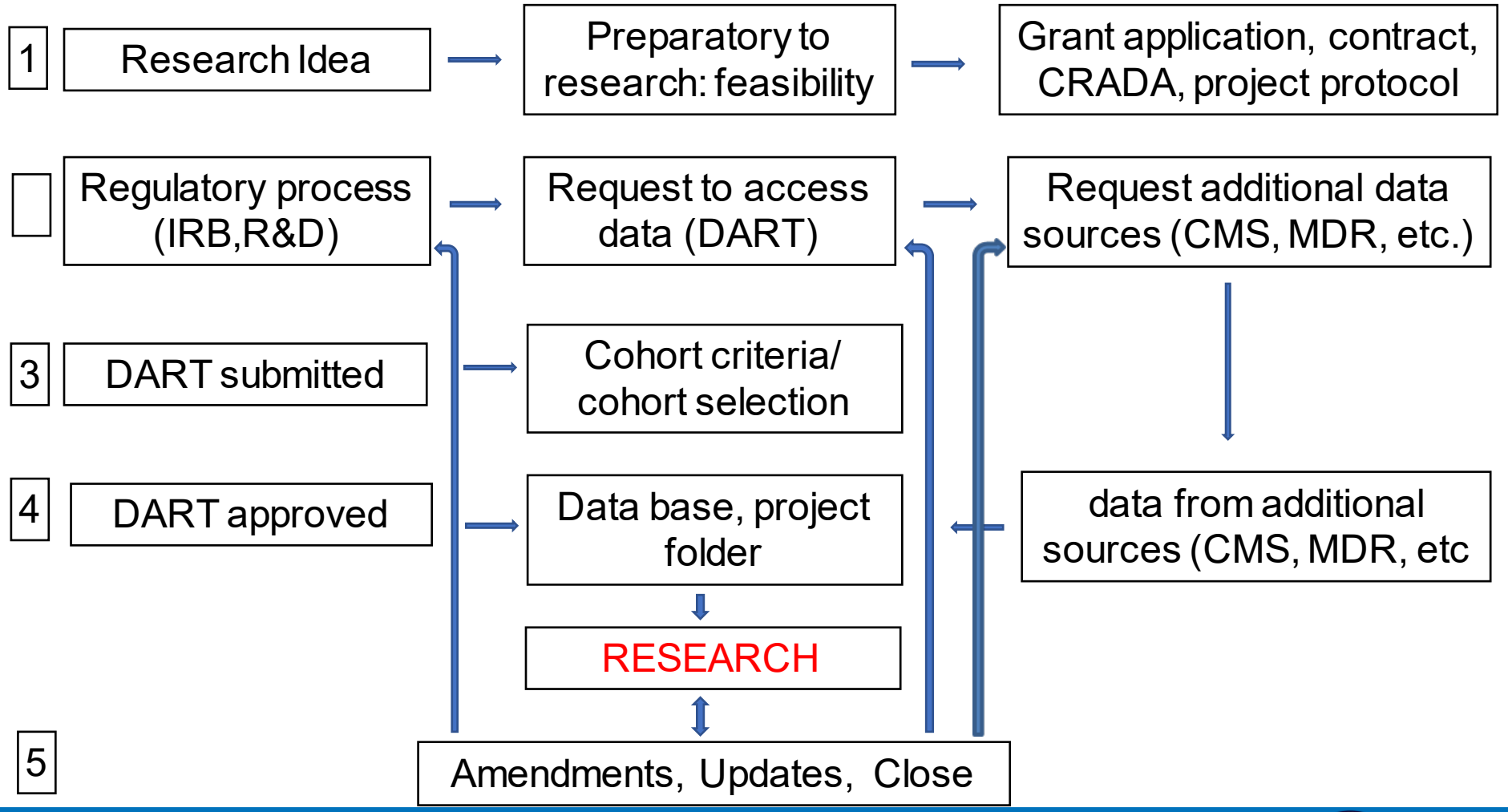
VA | U.S. Department of Veterans Affairs

Session Outline

- Research projects life cycle and resources
  - Preparatory to Research
  - Regulatory requirements
  - The Data Access Request Tracker
- VINCI Data
  - Skills Required
  - Data sources and domains
  - Cohort
  - Time range
  - CDW structure

"There are no unrealistic goals, only unrealistic timelines."
-Brian Tracy

Research Project Life Cycle



Preparatory to Research: Data Needs Assessment

VINCI Services supports VA research by assisting customers with Data Needs questions (DNA) relating to the VA Corporate Data Warehouse (CDW) and other VA data sources.

These questions range from

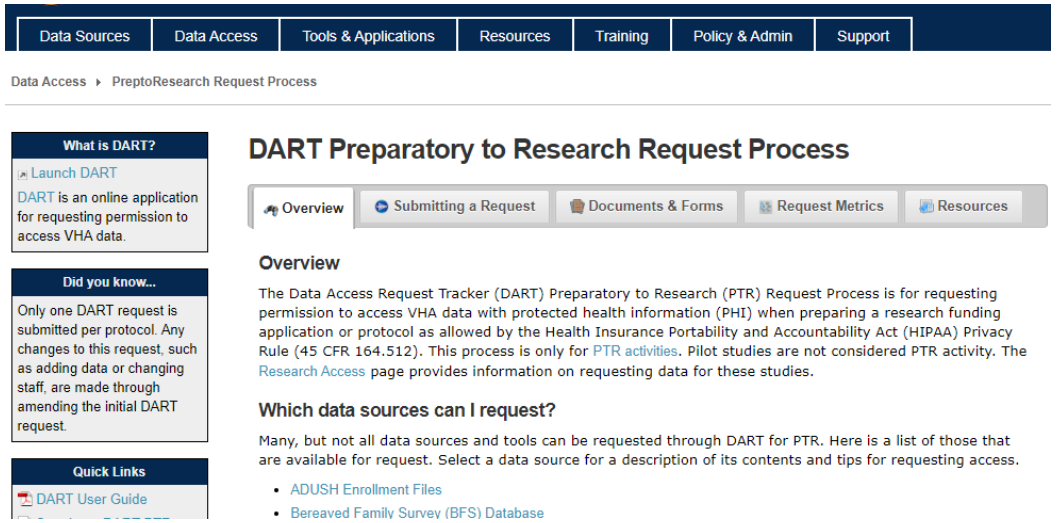
- ✓ data availability
- ✓ data location
- ✓ data definitions
- ✓ assistance with completing the DART application
- ✓ assistance with completing DART forms including CDW-domain checklist,
- ✓ assistance with IRB protocol preparation,
- ✓ etc.

Preparatory to Research: Feasibility

- Can be conducted by VINCI Services team (no DART) Concierge Services (va.gov)
- Feasibilities provide aggregate counts for researchers during the preparatory-to-research phase. The deliverables result in summarized counts to assist in grant and IRB applications.
- Feasibilities help to understand cohort size and structure
- Feasibilities help to understand impact of each criteria to cohort size and structure
- Feasibilities result cannot be used for any publication
- No regulatory documents are required

Preparatory to Research

- **Preparatory to research** Can be conducted by research team **(**Feasibility project (cohort size and structure; result cannot be published)
- SQL skills
- DART Preparatory to Research Request Process
- http://vaww.vhadataportal.med.va.gov/DataAccess/PreptoResearchRequestProcess.aspx#SubmitRequest

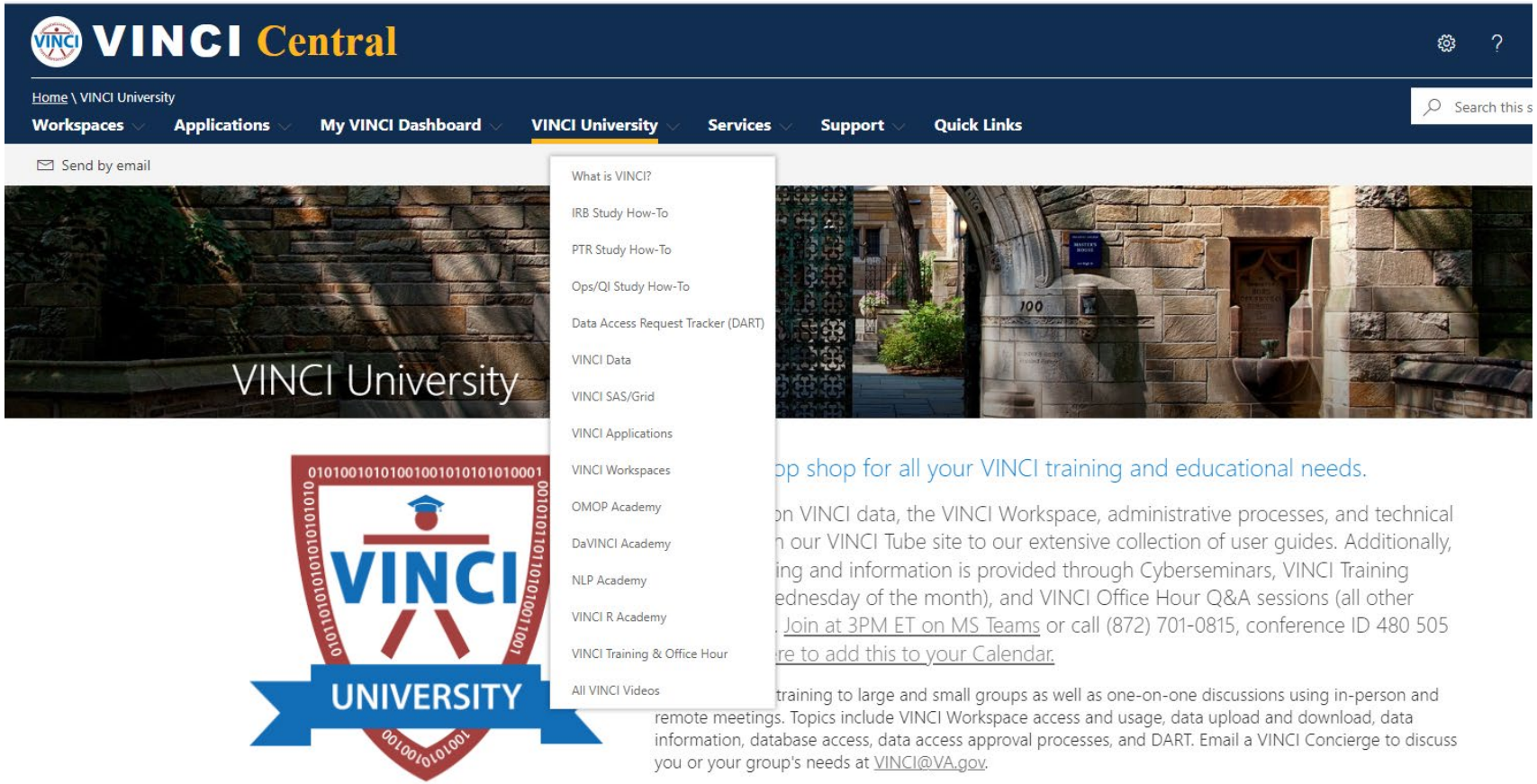


Regulatory: IRB/R&D

- **Research projects**
- IRB and R&D
- **DART (**The Data Access Request Tracker)/ Request list of patients from VINCI Services for the prospective projects/ Request access for the data for retrospective project
- http://vaww.vhadataportal.med.va.gov/DataAccess/DARTRequestProcess.aspx#SubmitRequest
- Forms for the additional data sources

Research Project: VINCI for IRB

[VINCI Central - Home (va.gov)]



Regulatory: IRB/ R&D

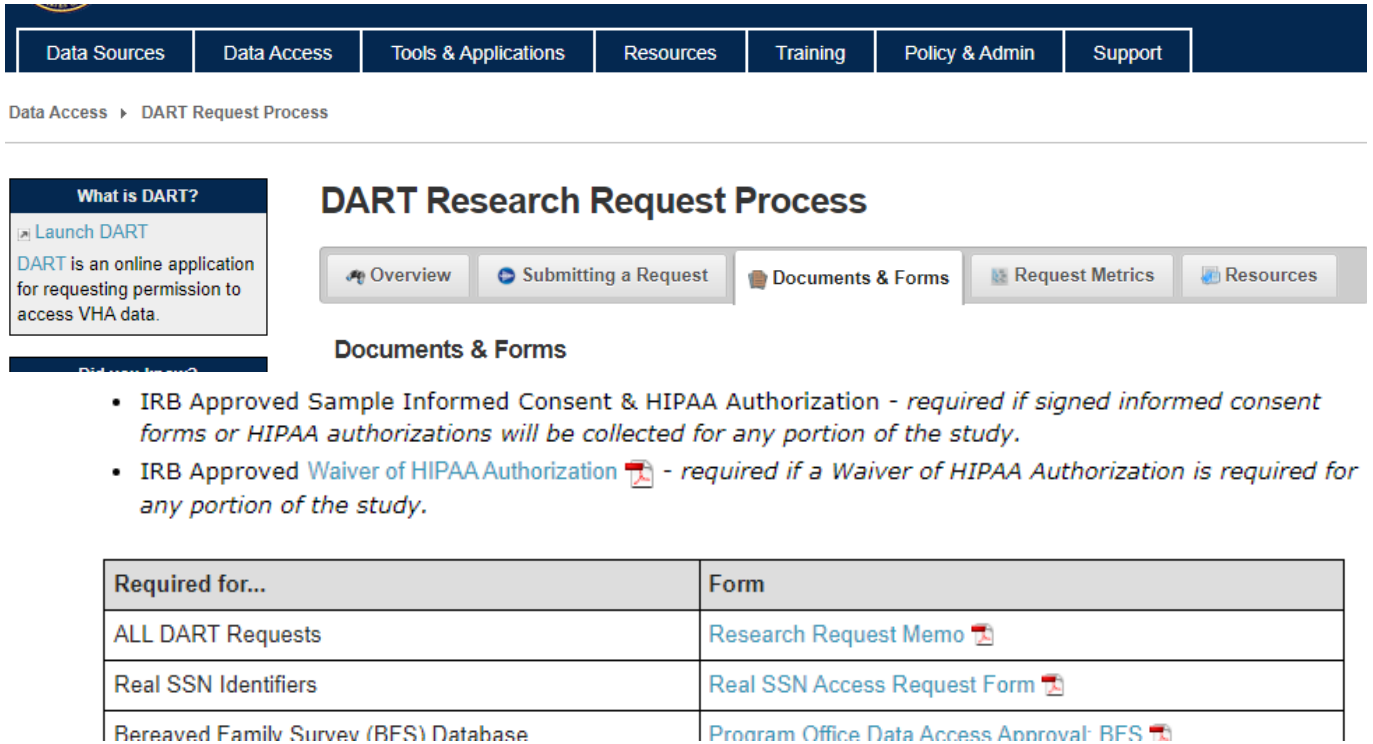
Required Information in Project Protocol

- VINCI description for IRB [IRB Study How-To (va.gov)]
  3. **As part of your IRB package you are required to state where data will be stored for analysis. If you are planning to use VINCI for data storage and analysis we provide the following environment and security descriptions for use with IRB packages.
     - VINCI Description for IRBs
     - VINCI Information Security Description
- Real SSN/ Scrambled SSN should be requested (CAPRI/JLV, IVC data, USVETS, non-VINCI data sources, other)
- If project will need **CAPRI/JLV access**, it should be described in the protocol/IRB protocol and HIPAA Waiver ("Access to patients records on national level will be conducting using CAPRI/JVL applications")
- If project required cause of death analysis using **MDR data**, it should be described in IRB protocol in study procedure section Mortality Data (va.gov)
- If project required **Medicaid /Medicare analysis using CMS data**, it should be described in IRB protocol in study procedure section VA/CMS Home
- If targeted population when using 7332 diagnosis (drug abuse, alcohol abuse, HIV, sickle cell anemia) the following statement should be included in IRB Protocol & HIPAA: The purpose of the data is to conduct scientific research and that no personnel involved in the study may identify, directly or indirectly, any individual patient or subject in any report of such research or otherwise disclose patient or subject identities in any manner.

  https://www.networkforphl.org/wp-content/uploads/2019/01/Snapshot-VA-Confidentiality-of-Sensitive-Records.pdf
- Storage location
- Inclusion/ exclusion criteria
- Waiver of HIPAA Authorization

Research Project: HIPAA Waiver

[DART Research Request Process (va.gov)]



Regulatory: IRB/ R&D

Inclusion/exclusion criteria

Ethics Principles for Access to and Use of Veteran Data

- Principle 7: All parties are obligated to ensure data security, quality and integrity of Veteran data.

All parties who send, receive, or use Veteran data must ensure data security, quality and integrity. In other words, that the data remain secure; accurate; complete; and representative of the data quality, meaning, and integrity when it was received or accessed from VA. Access to data should be limited to the minimum amount needed to accomplish the stated purpose, and should be terminated when no longer required. Data that are not necessary to accomplish the purpose for which it was obtained should not be retained longer than legally required. Transparency about breaches in data security, quality or integrity is also essential to promote trust and minimize impacts to Veterans.

Ethics Principles for Access to and Use of Veteran Data - DigitalVA

Regulatory: IRB/ R&D

Inclusion/exclusion criteria

- Location (nationwide or some specific locations HCS/MC/VISN etc.)
- Time frame
- Age/gender/other limitation
- Cohort definition with justification

Required Documents

- Research Study Institutional Review Board (IRB) Approval Letter
- Sample Informed Consent (for the prospective project)
- Research and Development (RD) Committee Approval Letter
- Research Protocol
- IRB Approval Waiver of HIPAA-Compliant Authorization
- DUA (if needed)

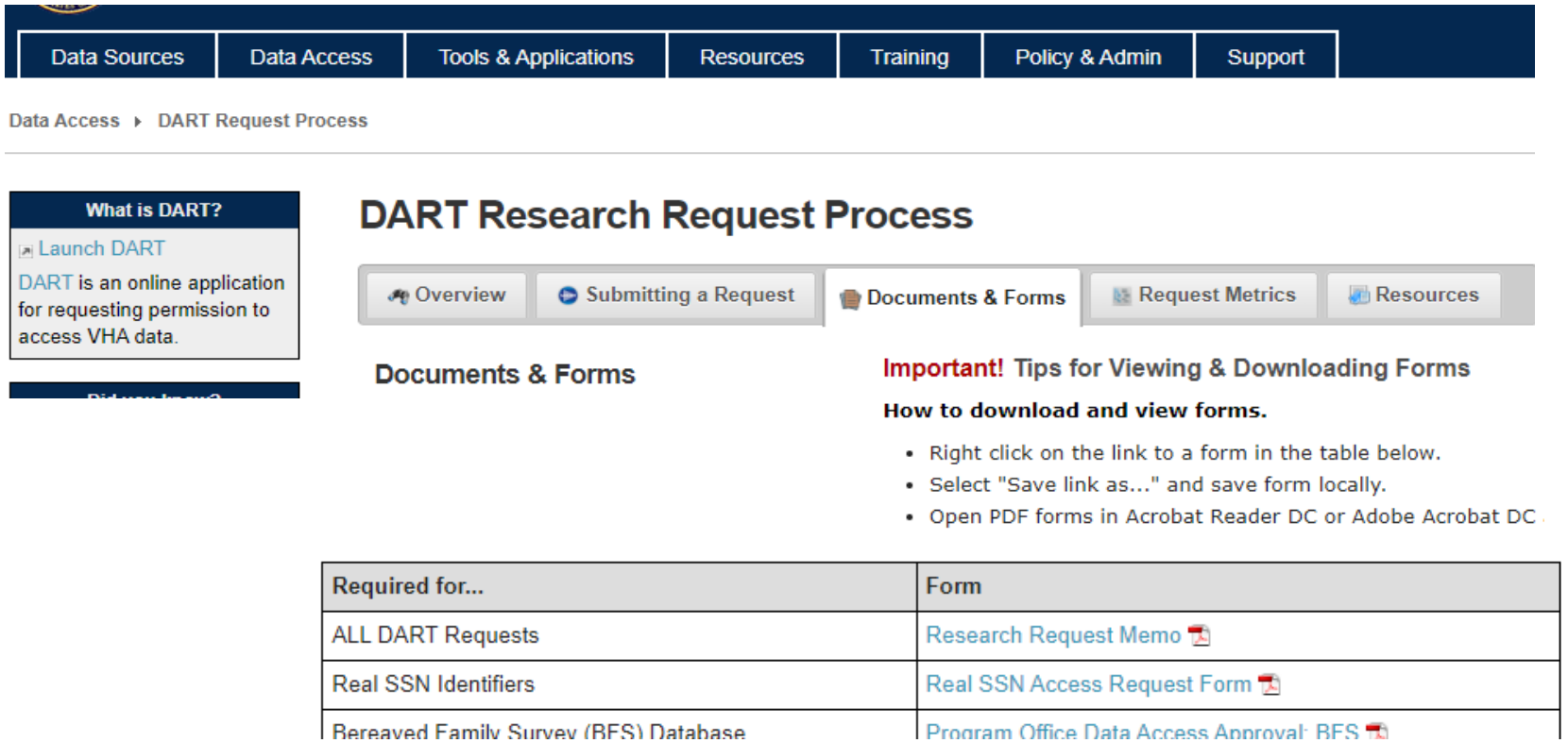
- *Research Request Memo**
- *Request to Real SSN Access (if needed – specific data sources (CAPRI, IVC data, chart review))**
- *CDW-Domain Checklist**
- *Millennium data check list**
- *Data source specific documents* (if needed – CAPRI/JLV, VASQUIP, etc.)*

* Blank documents can be found at [DART Research Request Process (va.gov)]

DART

The Data Access Request Tracker (DART)

http://vaww.vhadataportal.med.va.gov/DataAccess/DARTRequestProcess.aspx#SubmitRequest



DART Initiated Consultation

The **Data Access Request Tracker (DART)** Research Request Process is application for requesting permission to access VHA data for the IRB - approved research project

We strongly encourage team members to reach out to vinci@va.gov and request a Data Needs Assessment (DNA)/DART consultation before submitting completed DART application.

Consultation:

- **Discuss steps for the research projects using VINCI data**
- **VINCI data base: data availability and limitation**
- **Help with selection of the data sources needed for the research study**

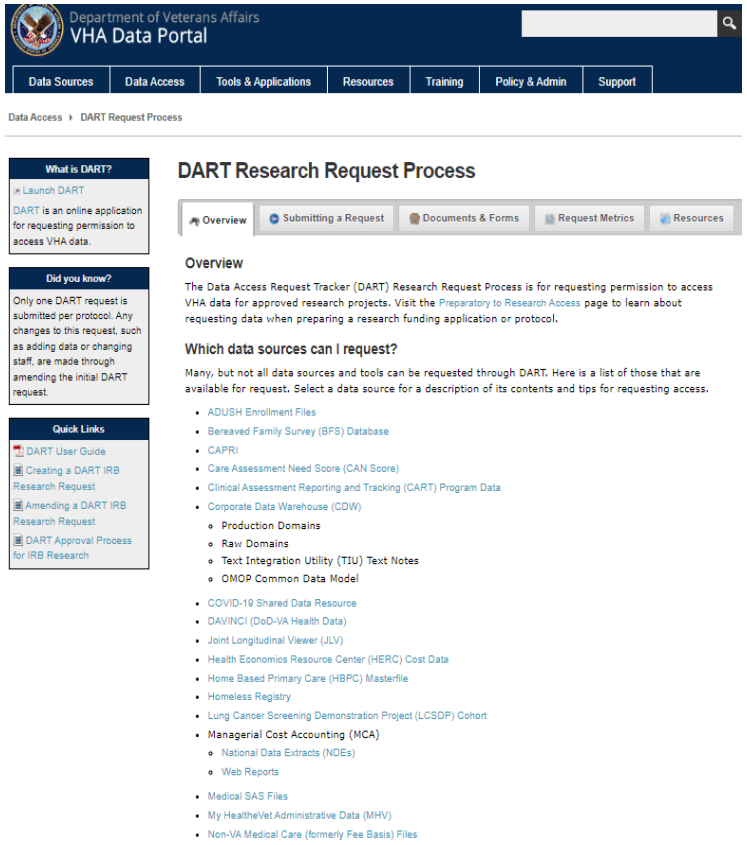
This will help ensure that the selections made on the DART Data Sources screen and CDW Checklist will result in the creation of a comprehensive dataset for the study.

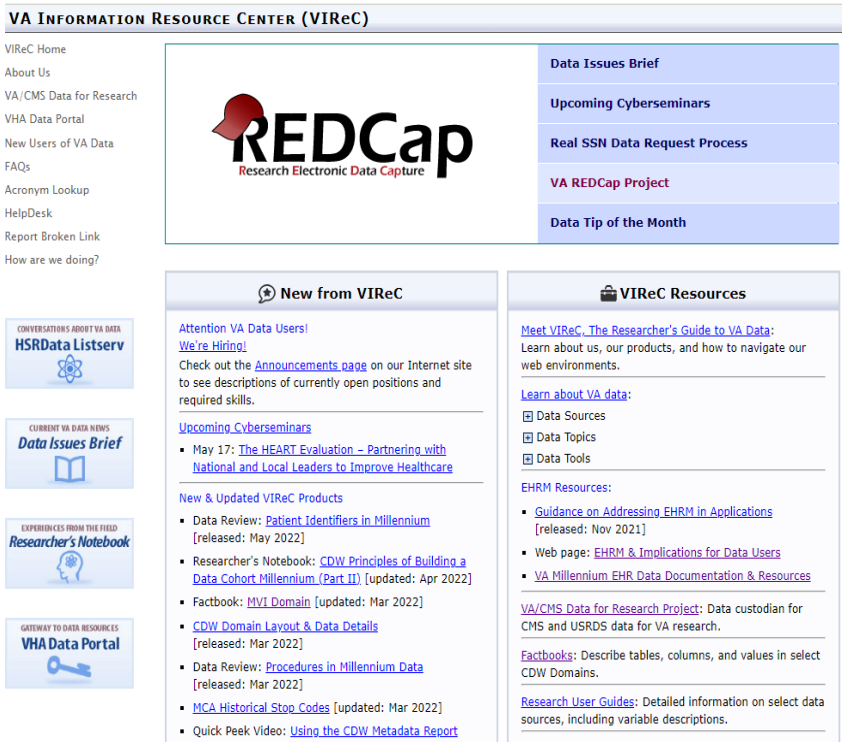
DART: Data Sources (More on This Later)

http://vaww.vhadataportal.med.va.gov/DataAccess/DARTRequestProcess.aspx#Overview

VINCI Data (va.gov)

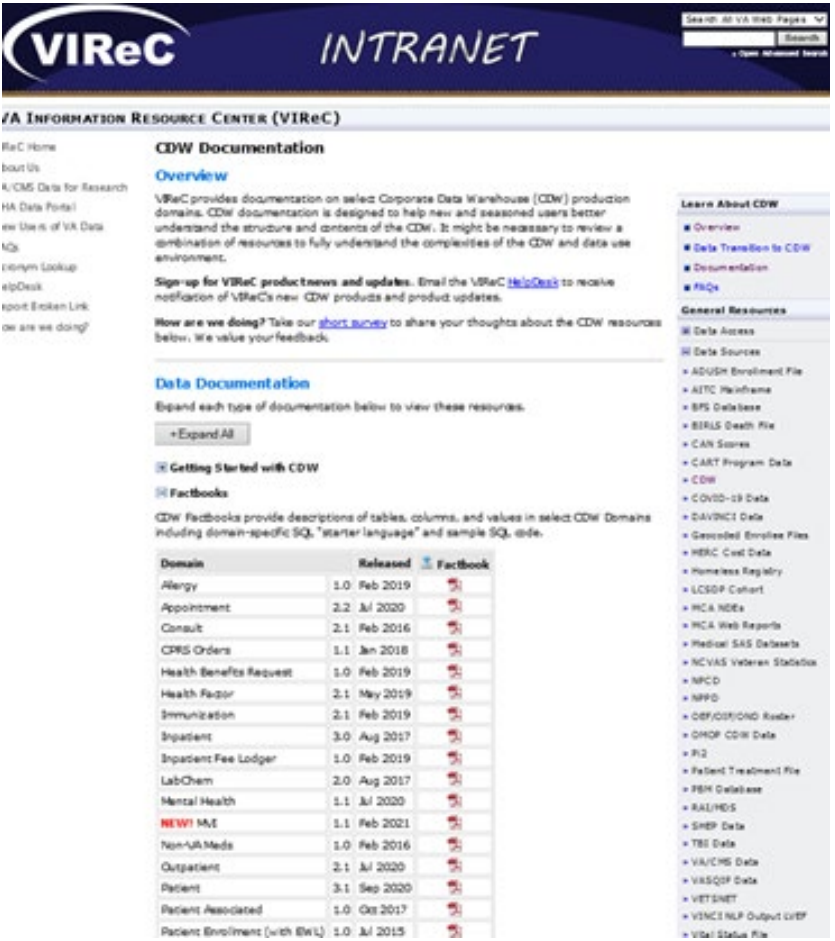
https://vaww.virec.research.va.gov/



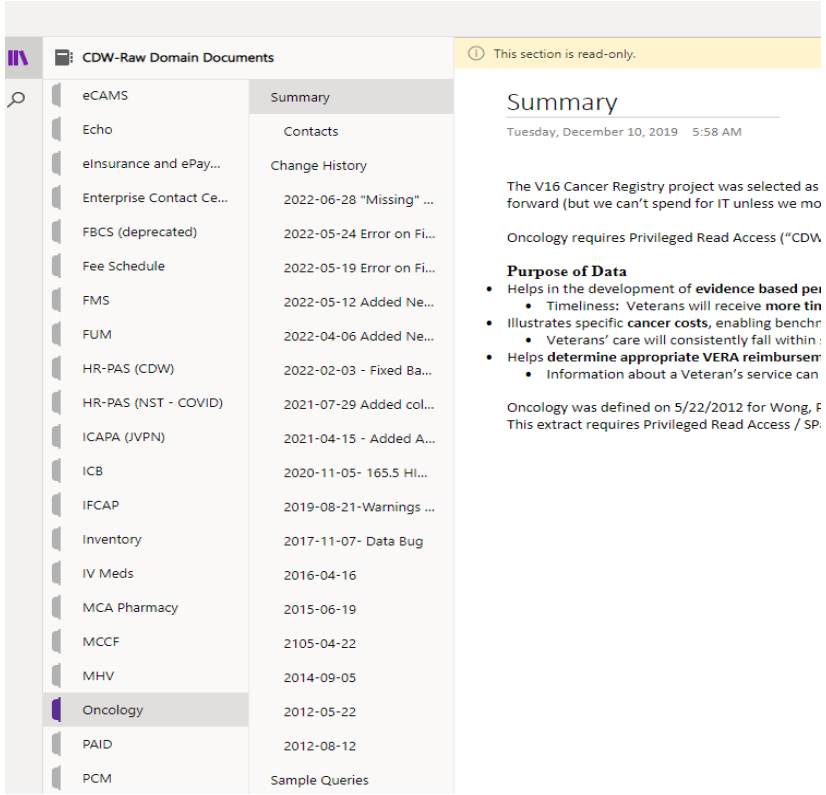


DART: CDW domain checklist

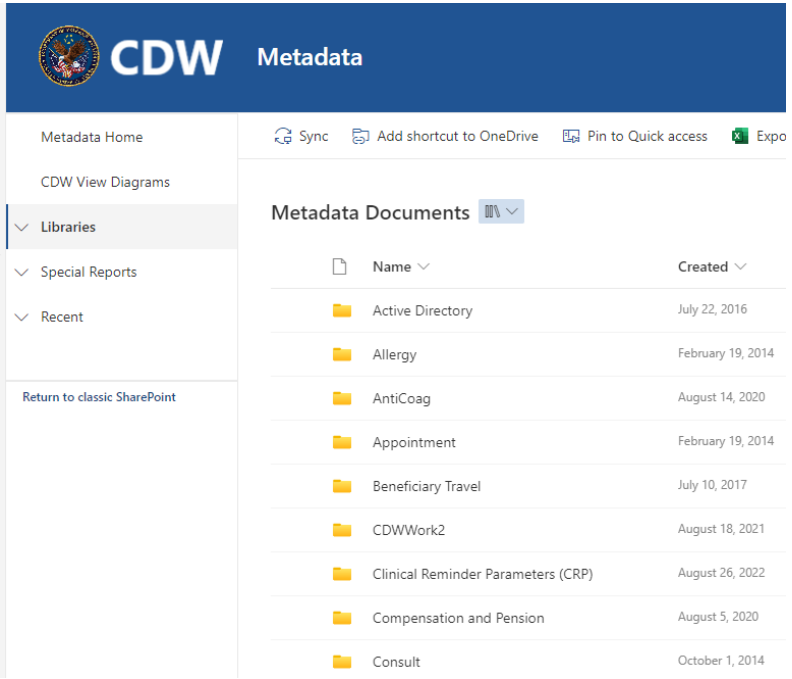
https://vaww.virec.research.va.gov/CDW/Documentation.htm#Factbooks



CDW-Raw Domain Documents (sharepoint.com)



Metadata - Metadata Documents - All Documents (sharepoint.com)



DART: Submission and Cohort Creation

Your research project with DART tracking number 2022-07-003-D has been submitted.

If your study requires a cohort derived from CDW/VINCI data sources, the next step is to provide a broad study cohort criteria that will be used to filter the VA patient population. Your preliminary cohort can be defined by specific diagnoses (ICD-codes), care settings (Inpatient/Outpatient), procedures, time period, etc.

Researchers will have two options:

- The first option (New and/or less experienced VINCI users may prefer this approach): while your DART application is under review is to work with our consultation team (VINCIServices@va.gov) to review your cohort inclusion and exclusion criteria. It allows more time for cohort definition refinement if needed prior to DART approval. This consultation allows for a wide-ranging review of various data sources and broad phenotype definitions to provide a comprehensive cohort that meets your study cohort needs. Please, send inclusion/exclusion criteria to VINCIServices@va.gov

OR

- The second option (More experienced VINCI users may prefer this approach as they may already be familiar with CDW/VINCI data and this process) : once your DART application is approved and correspondence site is created, is to work with the VINCI provisioning data manager directly and provide the cohort definition through the project correspondence site

Examples of **a cohort inclusion/exclusion criterion**, may include two or more of the following:

1. Location (nationwide or specific stations)
2. Time frame for each criteria
3. Diagnoses (please provide all relevant ICD 9 and/or 10 codes)
4. Procedures (please provide all relevant ICD-procedure and/or CPT codes)
5. Medication (please provide all generic and brand names)
6. Lab test (topography, units, range)
7. Type of project: Prospective or Retrospective
8. Other criteria

DART: Submission and Cohort Creation

Inclusion/exclusion criteria for the cohort = Inclusion/exclusion criteria IRB AND DART data sources

- Location (nationwide or some specific location)
- Time frame
- Age/gender/other limitation
- Diagnosis (ICD9*/10 codes (*VA uses ICD10 codes since October 2015. If time frame prior to 2015, please, provide both))
- Procedures (CPT codes or/and ICD-PC)
- Lab test. (topography, units, range)
- Medications (list of all generic and brand names, strength and route of administration)

DART Approved – Now What?

- “Where’s my flat file?” – What skills you need to work with VINCI Data
- Project data provisioned – What you get and how to access it
- The cohort – Who are you studying?
- The data – sources, domains and additional considerations

These slides will be more about managing the project than about managing the data

What Skills Do You Need to Work With VINCI Data?

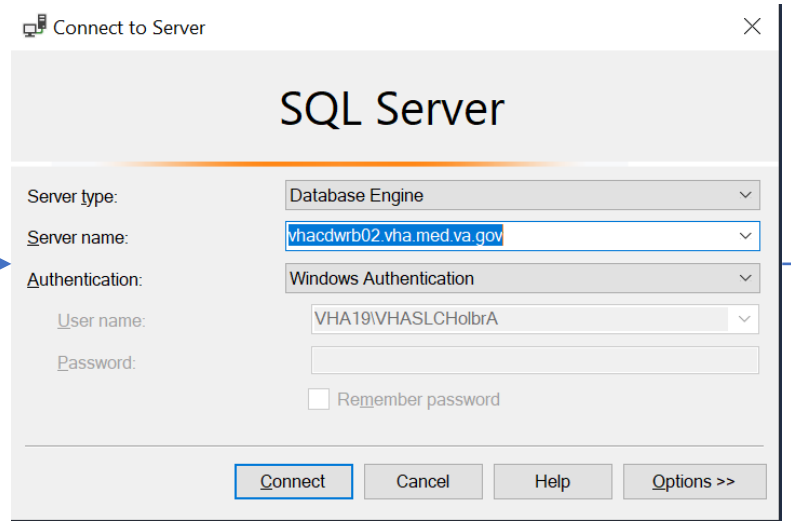
- Data Skills
  - VINCI data is provisioned in a normalized SQL database – you may have millions of rows and thousands of columns across hundreds of tables
  - You will need to communicate with the server in the language of SQL – this can be done directly in SSMS (SQL Server Management Studio) or as a pass-through from one of the analysis platforms below
- Analysis Skills
  - You have options! VINCI Standard Workspace Applications
    - SAS
    - R
    - Stata
    - MATLAB

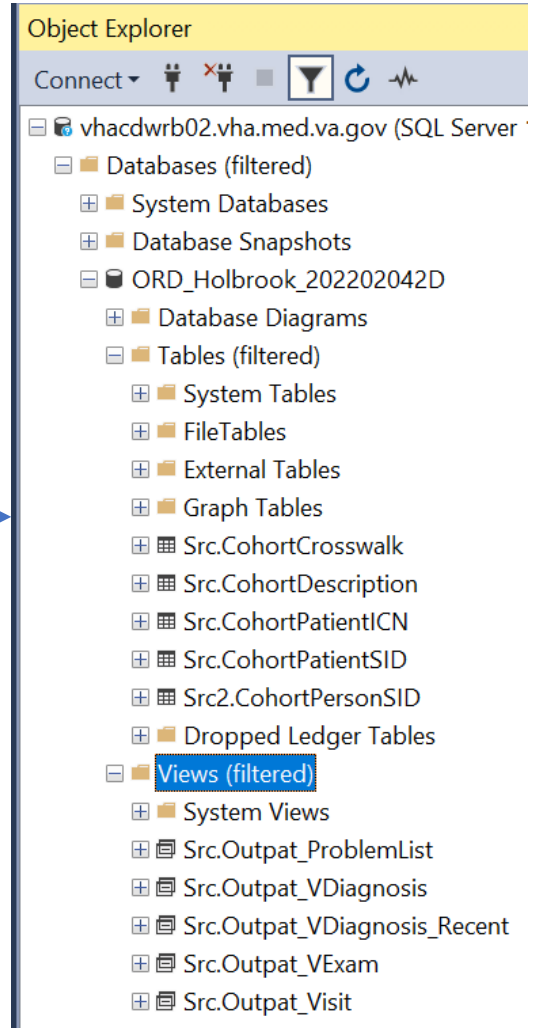
Accessing Your Data

- Log in to the VINCI Workspace to access your project folders and applications
- Everyone on the DART will automatically have access to the study folders and data

SSMS Example

SSMS from the start menu





SQL Pass-through Example (SAS)

```
PROC SQL ;
CONNECT TO OLEDB AS CDW1 ( DATASOURCE="VHACDWRB02.VHA.MED.VA.GOV"  PROVIDER=SQL_Driver
READBUFF=5000  CURSOR_TYPE=STATIC  UTILCONN_TRANSIENT=YES  DEFER=YES  CONNECTION=GLOBAL
AUTOCOMMIT=YES      PROPERTIES=( 'INITIAL CATALOG'=ORD_Holbrook_202202042D  'INTEGRATED
SECURITY'=SSPI )   ) ;
```

Using this pass-through method, you can write T-SQL to be executed on the server.

Source: "Best Practices in Accessing SQL Server Data with SAS" on VINCI SAS/Grid (va.gov)

Project Data

- VINCI provisions study data, capturing all records:
  - from the requested data sources and domains
  - for the cohort of interest
  - within the specified time range (optional)
  - maintaining the source data structures

Requested Data Sources and Domains

- Most data come from CDW Production
- Some examples of non-CDW data:
  - CMS
  - Death Ascertainment File
  - ADUSH Enrollment Files
  - CAN Scores
  - DoD Data
  - PAID Data

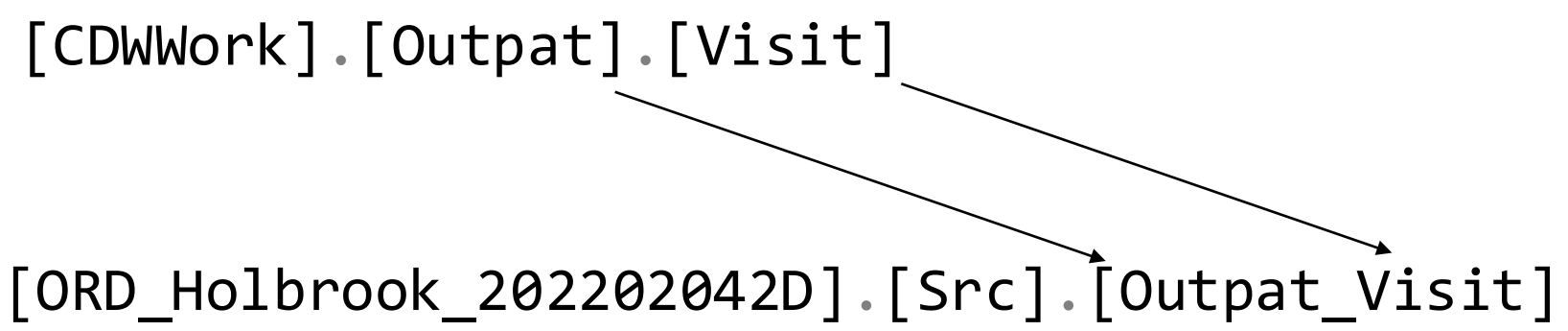
Select data documentation links

- Data Sources (va.gov) (VHA Data Portal)
- Working with VA Data (VIReC)
- Corporate Data Warehouse (OIT CDW)
- Data Sources (va.gov) (VINCI)

For lots more data links see this recent Cyberseminar from Brian Robison: Navigating CDW Data (va.gov)

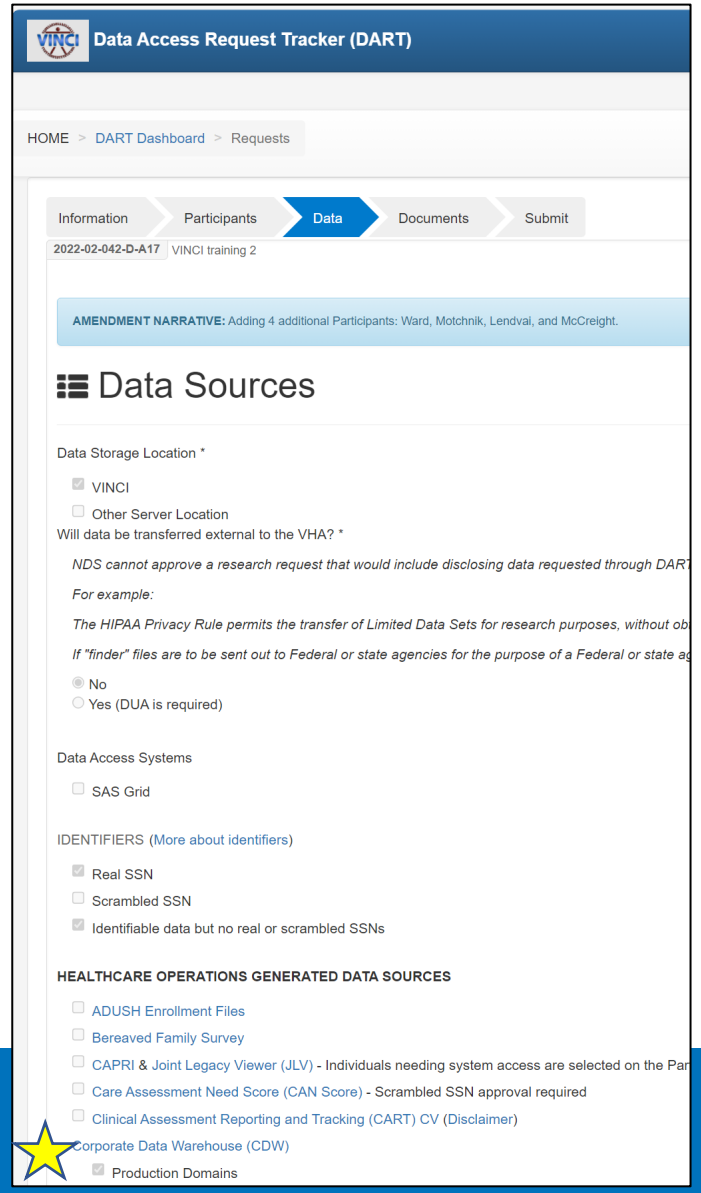
Data Domain/Source Naming

- Data from a particular domain or source typically live in one source schema
- In a research database, all provisioned data is in the src schema
- The original schema is prepended to the table name with an underscore
- Note that data documentation and metadata will refer to the original object names!



Where the Data Come From

DART Data Sources Screen

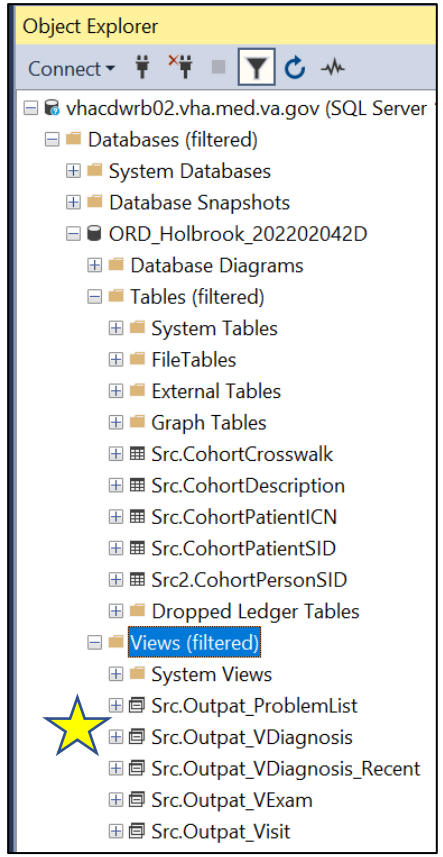


CDW DOMAIN CHECKLIST

CDW Production

- ☐ Allergy
- ☐ AntiCoag (Anticoagulation)
- ☒ Appointment
- ☐ Beneficiary Travel
- ☒ Consult
- ☒ CPRS Orders
- ☐ Dental
- ☐ Emergency Dept. Int. Software (EDIS)
- ☒ Enrollment (ADR)
- ☐ Health Benefits Request
- ☒ Health Factors
- ☐ Home Telehealth
- ☐ Immunization
- ☒ Inpatient
- ☐ Integrated Billing
- ☐ Lab Accession
- ☒ Lab Chem
- ☐ Lab Microbiology
- ☐ Lab Order
- ☒ Master Veteran Index (Veteran)
- ☐ Master Veteran Index (SVeteran)
- ☒ Mental Health Assessment
- ☒ Non-VA Meds
- ☒ Outpatient
- ☐ Pathology
- ☒ Patient
- ☐ Patient Associated
- ☐ Patient Enrollment
- ☐ Patient Insurance
- ☐ Patient Record Flag
- ☐ PCMM (Primary Care Management Module)
- ☒ Pharmacy BCMA (Bar Code Medication Administration)
- ☒ Pharmacy IV
- ☒ Pharmacy Outpatient
- ☒ Pharmacy Patient
- ☒ Pharmacy Unit Dose
- ☐ Program Integrity Tool (PIT)
- ☐ Prosthetics
- ☒ Purchased Care
- ☒ Radiology
- ☐ Recall Reminders
- ☒ Reengineered Primary Care Management Module (RPCMM)
- ☐ SPatient
- ☒ Staff
- ☐ SStaff
- ☐ Surgery PRE, INTRA, and POST
- ☐ VistA Compensation & Pension
- ☐ VistA Waitlist
- ☒ Vital Signs
- ☒ Women's Health

SQL Server Views



The Names Aren't the Same!

- Did you notice? "Outpatient" vs "Outpat"
- Most of these domain/source to schema name changes are obvious, but some are not.
- Raw and external sources often have "VINCI_"
- If you can't find one, ask about it in VINCI's weekly office hour or you can ask your data manager on the correspondence site

Some example name changes:

- PAID Data -> VINCI_PAID
- ADUSH Enrollment Files -> VINCI_ADUSH
- IVC Consolidated Data Sets -> VINCI_IVC_CDS
- Managerial Cost Accounting (MCA), National Data Extracts (NDEs) -> CDWWork_DSS 🤔

Cohort of Interest

- VINCI will provision data based on one or more study cohorts that you define
- You can have multiple cohorts
  - VINCI-provisioned cohorts (annoying for overlapping cohorts)
  - One big cohort from which you choose cohorts (best for complicated selection logic)
- Not too big
  - The larger your cohort, the more difficult and slow your queries/processing will be
  - Ethical principal 7
- Not too small
  - You can filter down to a smaller cohort yourself based on any data from any requested source/domain
  - You cannot see any data for anyone not in your study cohort – remember to consider controls

Where does the cohort come from?

Criteria Spreadsheet

Date of Query: 10/19/2021

Criteria Definition	Care Setting	Location	Time range	Query Definition	Domains/Data Sources Used:	Number of distinct patients remaining
Inclusion criteria						
Number of Veterans seen outpatient, inpatient, fee basis nationwide at period of time October 1, 2006 -present	outpatient, inpatient, fee basis	nationwide	October 1, 2006 -present	Number of Veterans, including those employed by the VA (see Additional Information for exclusion criteria) seen outpatient, inpatient, or fee basis nationwide from 10/01/2006 up to but not including the day of query.	Outpatient, Inpatient, Purchased Care, Patient	12,325,216
Number of Veteran nationwide with at least one stay in a VA CLC (nursing home)* October 1, 2006 -present	inpatient, fee basis			Number of patients from row 7 who had at least one stay inpatient in a VA CLC (see Ward Locations used in query) or one fee basis nursing home stay (see Fee Programs, yellow highlighted programs used in query), nationwide, from 10/01/2006 up to but not including the day of query.	Inpatient, Purchased Care	561,586

*1. Specialty like %nh%
*2. BedSection equal to 'CLC', 'NH', or 'NHCU'
*3. MedicalService equal to 'NHCU'

From communication with researcher

Usually more information on other tabs (CPT or ICD codes used, drug names, LOINC codes, etc.)

Data sources we used

Final cohort number should approximately match patient count in provisioned cohort and feasibility

Limitation:

Cohort has been created based on DART application Data Sources screen and CDW Domain Checklist. Cohort size may be changed and/or accuracy may be increased by considering the additional data sources below.

Enrollment ADR (additional veteran flag)

Data sources we couldn't use

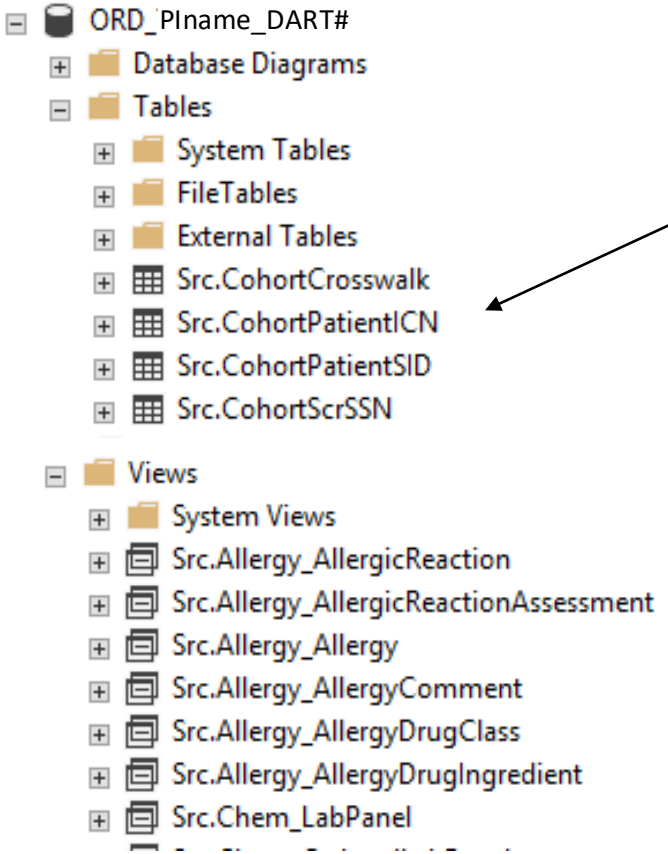
Time Range

- The time range used to define the cohort is separate from the time range used for provisioning
  - For the cohort: How do we decide who gets included? E.g. Patients with a knee procedure in CY2022
  - For the provisioning: How do we decide which records to put in your database? E.g. Records from 2012 - present
- Same not-too-big, not-too-small consideration as for the cohort

Data Structures

- All fact table views provisioned from CDW Production have had a name change from CDWWork (the “true source” containing all VHA data) to your study database, which is a subset of the “true source”:
  - CDWWork.Schema.Table -> ProjectDB.Src.Schema_Table
  - Example: CDWWork.Outpat.Visit -> ORD_Holbrook202202042D.Src.Outpat_Visit
- Besides that, all CDW metadata and best practices apply!
  - VINCI:CDW Meta Data Search SOP, VINCI - VA Phenomics Library
  - Data Services & Field Support (sharepoint.com)
    - Six Simple Steps
    - CDW Best Practices

Cohort Creation vs Data Provisioning



- The cohort
  - Created from methodology shown previously.
  - Static, unless you request otherwise.

- The provisioned views
  - Created from your cohort and time range.
  - Contain all records for those patients and time range.
  - Dynamic…sort of.

Tip: Request Cohort Updates If You Need Them

- If you have requested data up to present, then new records will appear over time…
- But only for your cohort!
  - Because remember that the cohort itself was static.
- And only for stations at which your cohort was seen at the time of cohort creation

VINCI checks provisioned data sources for new records **for your cohort's PatientSIDs*** daily.

*PatientSIDs are a station-level patient identifier. One patient will have different PatientSIDs at each station they've visited

Examples

- Patient has been seen in two stations at time of ORD creation. Study window is "up to present."
  - Both PatientSIDs are captured.
  - All outpatient diagnoses are captured from those two stations, with new ones appearing daily, as the patient completes encounters.
  - If the patient goes to a VA where they have never been seen before, a new PatientSID will be created in the CDW.
    - It will not appear in your cohort, because the cohort table is static.
- A different patient never got the diagnosis of interest, and so was not captured in the cohort.
  - If the patient gets a visit today with the diagnosis of interest, they will not appear in the cohort.

You Can Request Cohorts Be Updated

- Use the correspondence site, just like for other communications.
- This will “fix” both of the issues from the previous slide.
- Alternatively, you can request static data on the rb01 server

Other Tips

- Remember that your project protocol/IRB approval should match what you request in the DART
- If you do need to update your DART for a new data source, remember to post to the correspondence site after approval

Final Tips

- We can help!
  - Feasibilities/DNAs are valuable.
  - Join office hours: VINCI Training & Office Hour (va.gov)
  - Ask questions or request DNAs/Feasibilities at VINCI@VA.gov

Acknowledgements

- VINCI Data Services and VINCI IT
- SAS/Grid Admins
- VIReC
- CDW Data Services and Field Support

QUESTIONS?

THANK YOU!