Brian Robinson: All right, thank you, Rob. So as Rob mentioned, I am Brian Robinson. I am with the VINCI Data Services Team, and today I'll be giving a presentation on navigating CDW data. And so just to check Rob, are my slides coming through okay?

Rob: Yes.

Brian Robinson: All right, great. So my agenda today is to provide an overview of VA data sources, to walk through some tips on finding research variables of interest, to explain and understanding of patient identifiers that are available in VA data, and to talk a little bit of SQL indexing. And the VA data is vast. There is a ton in here, and so, of necessity, this presentation is going to provide a very high-level overview of all of these different domains. Many of these domains could really be a presentation of themselves, and so this will be again a very high-level run through. And these slides will be available. I am providing a ton of links to documentation and other resources so hopefully you can take this and refer back to it and dive deeper in those domains that you're interested in.

So starting with the VA data sources, they can more or less be divided into kind of three categories. So we have data that comes in from the CDW. We have our non-CDW data, and then we have other data, and so I'll cover each of these in turn. So starting with CDW data, this consists of data that comes from CDW production and CDW raw data. The data will reside in either the CDW work database on VINCI servers or provision to your ORD databases. This data has been standardized. It's been reviewed by clinical subject matter experts. The data is categorized into Dim or reference data and fact data. These tables may combine multiple vista files into a single view so it's not a one to one coming in from VistA necessarily.

There are SIDs that are added by CDW, which makes joining across tables easier. There are indexes that are added by CDW, and there is documentation that exists -- it's not great -- it's not horrible -- and some of that documentation includes the CDW support site, Metadata report, and the Meta schema within CDW work itself. As I mentioned, first I'll talk about dimensions or Dim tables. So the dimension data do not contain any patient data. All the dimension tables are accessible to all studies. You don’t need any kind of approval or anything like that to get this data because it does not contain any patient data.

What it does contain is reference and lookup information, including information like ICD codes and diagnosis codes and the descriptions; procedure codes and descriptions and numbers; CPT procedure codes; drugs and drug classes; lab tests; specimen types; information about VistA sites and stations, visions, and regions; and other descriptive information about the terms and items that you may find in the VA healthcare data. This can be used to efficiently identify patients for cohorts that have conditions you're interested in and to get the data you're after.

Of note, the NDIM schema is used as a determination for any dimension table that is not actually sourced from a VistA system. All right, and so I'll run through the different domains we can find in CDW work. So first, we have the ADR enrollment. This contains data from the enrollment system administrative repository. This includes information about Veteran status, enrollment status and priority group, and it contains fact and Dim tables.

Rob: Brian, we have a question. Somebody is asking what is a SID?

Brian Robinson: So a SID is a station identifier, and I will cover that when we get to kind of the patient identifiers. It can also be a thing that is used as a primary and foreign key within the tables to allow for joining across the tables in a relational structure. So hopefully that answers that question. So we have the allergy domain, and this domain contains allergy and adverse drug reaction information. It includes information about hypersensitive reactions, including reactions to medications, foods, chemicals, and noxious stimuli, and it contains specific information about the type of reaction that occurred as well. And the fact tables and Dim tables are these.

The AntiCoag domain, so this domain contains information from a combination of VistA file 130 and the ORAM flowsheet through the anticoagulation management tool. Note that not all 130 facilities use this tool so data will be missing from those facilities. This domain includes data on international normalized ratio or INR, and it also includes data on the indication for the anticoagulation treatment. And it contains these two fact tables. The appointment domain contains information about appointments, including appointment dates, the purpose of the appointment; the status of the appointment; if it was canceled, the reason why, and the staff member who canceled it; and it uses these fact and Dim tables.

The BeneTravel domain contains information on records of travel payments to patients and patient caregivers, and it includes these two fact tables and the one Dim table. The consult domain contains information about requests for services, including clinical specialty care, procedures, and non-VA care, and it uses these fact and Dim tables. We have the Covid-19 shared data resource. So this contains information related to Covid-19 both in and outside of the VA. This is a collaborative effort, which utilizes multiple data sources and is a collaboration across multiple teams. It includes case definitions, concept definitions, VA data mappings, and other information all related to Covid-19. There are 33 fact tables and seven Dim tables in this domain.

The CPRS orders domain contains orders and requisitions made through the order entry option using any CPRS package. This can include things like labs, radiology orders, medications, etc., and it uses these fact tables. The dental domain, as you would guess, contains dental data. It includes records of dental exams, diagnoses, history, transactions, and procedures, including basic hygiene, restorative surgery, orthodontics, periodontics, and endodontics. There are 29 fact tables and 17 Dim tables in this domain. We have the emergency department integration software or EDIS, and this is the data on patients as they're tracked through the EDIS software. And it includes patient arrival, disposition, departure, and reason for visit, and it uses these fact and Dim tables. We have health factors.

The health factor domain contains information about over 3000 different types of factors, which may affect the health of specific patients. Some common health factors include tobacco use, alcohol use, and drug use. The usage of health factors will vary substantially by station. The health factors are often used with clinical reminders, and the health factor data is mostly free text. So it comes with a bit of work to be able to comb through that if you want to dive into this data, and it uses this fact table and Dim table.

The immunization domain contains records of individual patient immunizations. There are over 52 million immunization records in the immunization domain, and the records can contain links to outpatient visit records via the VisitSID. So there is that SID again. So again, that's the station identifier, and anything that has SID in it is a station identifier that is generally used as a primary or foreign key when you're joining across tables in the relational database. It uses this fact table and the Dim tables.

The inpatient domain is generated from the inpatient VistA treatment file number 45, which contains all of the information generated from admissions, treating specialty transfers, and PTF screen edits. It can include records of inpatient admission; discharge; surgical procedures performed; ICD and CPT procedures; the primary, secondary, third, fourth, etc. diagnoses included during the inpatient stay; and the ward location. The primary tables used in the inpatient domain are these.

The integrated billing domain contains data from bills generated on behalf of patients. This is generated from the VistA both bill claims 399 and accounts receivable 430. There are 36 fact tables and 31 Dim tables in this domain. The lab accession contains information on lab pathology data from VistA file number 68, and it uses these fact tables. The lab chemistry domain includes information on lab data for patients, and we recommend you use the PatientLabChem table as the recommended path to access LabChem data for most patient analysis projects because this table restricts the laboratory results to only those records that refer to human patients. And it will filter out any administrative, referral tests, and research test records. It uses these tables for the primary domain tables, and these are the primary Dim tables in this domain.

The mental health assessment domain contains data collected during administration of a specified instrument from the mental health test and survey file. It's given to a patient at a specific date and time. So this includes things like the PH9 and things of those nature so different mental health assessment instruments. Please note that some of the survey text fields in this domain can include hidden characters like tabs and carriage returns. So you can use the replace function, along with ASCII codes and SQL to identify and remove these. The microbiology domain contains data on report level and organism antibiotic susceptibility level records. This also includes etiologies involved and pathogenic microbiology. Much of this data is text based, so again it comes with the issues there, and it includes 15 fact tables and two Dim tables.

The Master Veteran Index or MVI is a single master index of all Veterans in the Veterans Health Administration. It acts to maintain and synchronize identities and demographic data for VA clients, Veterans, and beneficiaries. It includes PatientICN status to help assist with identifying deactivated PatientICN's, and there are four fact tables and 11 Dim tables in this domain.

The OMOP domain is the international common data model. It uses standard vocabularies. It standardizes the data allowing for data interchange and uses the OMOP standards of conditions, procedures, device exposure, drug exposure, drug dose, drug era, measurement, and observation. So it uses these defined table structures for each of the data domains, and it's in a person- and provider-centric model. So of note, the OMOP domain includes the data coming in from all the other CDW domains, and they are transformed into the common data model.

The outpatient domain contains outpatient visits, and the outpatient visit domain replaced the CDW encounter domain. Unlike the encounter tables, no records are filtered out of the outpatient domain, aside from the normal exclusion of records before October 1, 1999. It pulls in data from the VistA encounter and VistA 9000 related files, and the primary tables are these. The pathology domain provides data from several pathology VistA files. It includes information on autopsies, cytopathology, and surgical pathology, and it includes 69 fact tables and two Dim tables.

The patient enrollment domain originates in the VistA version of the VA enrollment data. So when a Veteran goes to a VA medical center to enroll in VA care, information pertaining to his or her enrollment is stored in the local VistA system and then is sent to the health eligibility center for verification or HEC.

The HEC enrollment data is considered the authoritative source on enrollment. Therefore, this VistA enrollment data is synchronized with the HEC data once a determination is made. There is a field in the CDW table for source of enrollment, which indicates whether or not a particular record has been synchronized. This domain contains information, such as enrollment status, dates of enrollment periods, severity of disability, service-connected disability information, military discharge status, type and number of benefits being paid, priority group, etc. And it uses the patient enrollment table.

The Pharmacy Barcode Medication Administration or BCMA domain contains medication administration, process for the inpatient setting. So this is medication administered in an inpatient setting. It includes dates and times the medication was ordered, delivered, and administered. It includes details on the medication. It includes variants or deviation from the administration processes, and it uses these primary tables. The pharmacy outpatient domain includes all outpatient pharmacy data, including administrative pharmacy services and info on patients receiving prescriptions, things like fill dates, process dates, the number of days' supply, information on refills and renewals. It also includes specific drug information like the drug name and the VA drug class. It uses these primary domain tables.

The Program Integrity Tool or PIT includes data on VA claims that are processed through the PIT. It includes but is not limited to choice claims. There are 40 fact tables and 27 Dim tables here. We have the purchased care authorized or fee domain. This includes data on reimbursement of non-VA medical care from an external entity. It includes the vendor, the reason for the visit, the dates of service, status of the reimbursement, the amount filled, the amount paid, type of payment. It includes patient information, such as diagnoses, insurance, admission and discharge and dispositions. There are 12 fact tables and 9 Dim tables in this domain.

The prosthetics domain contains data on prosthetics and prosthetic purchasing transactions. It's generated from VistA file 660, and it includes 27 fact tables and 22 Dim tables. The radiology domain includes administrative and clinical information for radiology exams, orders, and reports. It uses two different schemas to identify what is going on with them. So the Rad schema indicates that it is radiology fact tables, and the SPatientText schema indicates that it includes report text, notes, and written clinical histories. There are 20 fact tables and 21 Dim tables in this domain.

The Rx IV domain contains information on VA IV activity, including renewals, label tracking, and comments from providers. It is generated from VistA file 55 and includes seven fact tables and two Dim tables. We have the patient and SPatient domain. So this domain includes all patients that are followed by medical centers and/or outpatient clinics. It includes demographics, such as age, race, and ethnicity. It includes sensitive patient data, such as name, address, and real social security number. It includes military service information, like service-connected disability and period of service, and it uses these primary domain tables.

The staff and SStaff domain includes information on staff who are affiliated with the VA. The SStaff domain, much like the SPatient domain, indicates whether it includes secure PII, including things like name and social security number whereas the staff without the S in front, same as the patient without the S in front, will contain more basic non-PII information. And it uses these primary domain tables.

The surgery domain includes pre, intra, and post is data that is extracted from the surgery package. It is split vertically into these three conceptual groupings of pre, intra, post. The pre includes case scheduling and preparation, the intra includes data during the surgical case progression, and the post includes post-case data, provider notes, indications for the operation, and surgically-implanted prosthetics. So each view has a case identifier; a patient identifier; a date and time of the surgery, location, etc. so you can track them across the three conceptual groupings. It uses these primary tables.

The Text Integration Utility or TIU contains free-text documents, such as progress notes and consult reports. The TIU notes from VINCI are full-text indexed to allow for easier searching, and when you are searching through these free texts please utilize the table value function listed here for a full-text search. This table value function is provided in CDWWork and provided to all of our research databases. So when you're searching the full text of the TIU note, use this table value function for best results. The data is extracted from the VistA file TIU document 8925, and it includes these domain tables.

The vital signs domain contains vital sign data from all sites. It is extracted from the VistA file GMRV vital measurement 125. There is no unit of measurement with these vitals. However, it is assumed that height is measured in inches, weight is in pounds, respiration is in breaths per minute, pain is from 1-10, pulse is in beats per minute, temperature is in decreased Fahrenheit, etc. or the vital qualifiers, which are core things like the arm from which the blood pressure was taken. It uses the vital sign domain table and the vital type Dim table.

The women's health domain provides information from the women's health VistA files in the 790 range. So the information in this domain contains administrative and clinical information pertaining to care for female Veterans and includes these domain tables. Now, we have our CDW-raw data. So data that is considered from the CDW-raw is data that has gone through minimal processing. The Raw team's goal is to simply get it out of VistA and into SQL so swim at your own risk. They're raw for a reason. You must use the station and IEN for joins with some exceptions, and I'll cover this more in my identifiers. The data is not indexed with some exceptions. There is documentation, but you may be sorry you asked. Data in this domain includes the echocardiogram.

The echo domain contains echocardiogram procedures done on a patient. This does not have echo reports. However, some echocardiogram results are located in the TIU notes. The echo domain is a limited extract of the following VistA files among others. The oncology domain is a CDW-raw domain that does a limited extract of the VistA oncology module.

The oncology module automates the tumor registry and supports tumor registrars in abstracting cancer cases, following up on cancer patients, and producing the Hospital Annual Report. Then we have our non-CDW data. So non-CDW data usually is highly-refined datasets, which may be based upon VistA and other data sources. It's usually very clean. It tends to be de-normalized or flat and non-relational. The main patient identifier is typically the scrambled social security number but sometimes uses the patient ICN. The data here typically predates the concept of CDW. It tends to be intended for very specific business needs or in other words very focused. It tends to have some demographic information but usually not complete, and there is reasonable documentation for this data. So among this data, we have the PSSG.

The PSSG or Planning Systems Support Group is geocoded enrollee files that is a group of files that are quarterly datasets comprising detailed and extensive geographic information. This data contains one record per enrollee. So VINCI receives the SAS files, and the data is loaded in the SQL server for research access, and the PSSG data includes data on patient address, including things like drive times. The VASQIP domain or VA Surgical Quality and Improvement Program includes surgical outcome data, including morbidity and mortality. It includes data on the quality of processes and structures of surgical care.

Then we have other data. So this tends to be datasets that may be based upon CDW data but not formally produced or managed by CDW. Documentation can be hit and miss. It frequently contains patient ICN or scrambled social as the main patient identifier and indexing can be hit and miss. This includes the Care Assessment Need or CAN score. So the Care Assessment Need or CAN score reflects the estimated probability of hospital admission or death within a specified timeframe of 90 days or one year. The score is expressed as a percentile ranging from zero, the lowest risk, to 99, the highest risk, compared to other VA patients. The data used to generate this score is obtained from VHA administrative databases, and in the literature you'll find that it is considered very accurate for large groups of patients.

The DaVINCI domain includes a Department of Defense and VA-VINCI data integration and collaboration. So this is an interagency collaboration based on resource sharing. So data from the military health system and other service-related information has been shared with us from the Department of Defense and integrated into the VA data through the help of VINCI. It includes data from the DOD on healthcare events, historical beneficiary data, service-related information, and demographics.

The left ventricular ejection fraction or LVEF domain contains information on the LVEF measurement extracted using natural language processing. It measures the total amount of blood in the left ventricle that is pumped out with each heartbeat, and the measurement is given in the range value. So the LVEF dataset separates the lower and upper values of each range that a patient receives, and this data is available from VINCI upon request.

The Patient Aligned Care Team or PACT is the VHA's implementation of the patient-centered medical home model, which aims to improve health outcomes through team-based care and improved access and care management. It includes composite scores that were constructed for eight core domains, including access, continuity, care coordination, comprehensiveness, self-management support, patient-centered care and communication, shared decision making, and team-based care. The main outcomes and measures are patient satisfaction, rates of hospitalization and emergency department use, quality of care, and staff burnout.

All right, so moving on I'm going to talk a little bit about how you can find research variables of interest across all of these different VA data domains depending on what you're interested in. So I'm going to go through some tools you can use to help you. Commonly-used data sources when conducting research in the VA includes CDW production, CDW raw, the vital status, and OMOP domains. So there is documentation on CDW production across multiple sources, and all these links I include at the end in my references. So we have VINCI Central homepage.

We have the Meta data report via the CDW share point site. We have the Meta data viewer application via VINCI central. We have the Dim data viewer application via VINCI central. There is documentation from VIReC, including their fact books. There is the Meta view within CDW work databases, and there is the VHA data portal. So I've included just some screenshots and links to many of these. So first we have the BISL Meta data report tool. So you can see here we can go to this tool, and we can select through on either ideas or domains we're interested in.

So my example here I'm interested in the women's health domain. So I've selected that. You can see in the report here, and what it does is it displays for me all of the different views that are in this domain, and from here I could select these views individually and see what fields are in those domains, see the ER diagrams and their relationship between these views and what each field contains, a description of each field. So you can use this reporting tool to kind of dive around the CDW data and figure out each individual domain, what's in there, how it relates, and how you can navigate it.

We have the VINCI Meta data viewer from the VINCI central site. So what this does it is allows you to access the meta data so you can use key word searches, and it will tell you where you can find those things here. And so I've included this example. Let's say I'm interested in CPT codes. So where do those show up within VA data? So I search for CPT. It shows me the schemas it's matched so I've selected inpatient. I'm interested in inpatient procedures using CPT codes. It shows me that there is a view for inpatient CPT procedures so I've selected that, and down below you can see its showing me the fields that are available in that inpatient CPT procedure view.

We also have the VINCI Dim data viewer. So like the Meta data viewer, this one allows you to view the Dim data through the VINCI central site. So I'm interested in smoking so I've done a search for smoking. I've selected that I want smoking as a finding, and the category of the Dim I would like is to look at the health factors because I recall that the health factor domain includes smoking information. So you can see here that this shows me the different health factor terms that show up in the health factor domain related to smoking.

We have documentation from VIReC, and I've screen shotted their linked care from their page. So they have some excellent resources here, including getting started with CDW. You can see there I can dive into that. They have excellent fact books. They have the researchers notebook and other information on CDW data, including snapshots, domain layouts and descriptions, contents, and frequencies. You can use the Meta views within the VINCI workspace. So if you're in SQL server inside the VINCI workspace, you can see here that there are Meta views. So you can search them from within SQL here.

So the example I've used here is I'm searching for the view fields that might contain the term ICD-9. Let's say I'm interested in ICD-9 codes and where they may show up, and so I do this search for any views that contain a view field with the key term ICD-9. And you can see here I've found it across some different tables, including inpatient specialty transfer, outpatient VProcedure, inpatient surgical procedure and so on. The CDW raw documentation is provided here from the CDW raw share point. You can see here I can select various domains that are from CDW raw so I've selected oncology, and it's showing me a breakdown here displayed of the vista files and fields that are being pulled into the oncology domain.

We have the vital status files so this contains information on death dates and other information around mortality, and so I've included a link and a screenshot here to the VHA data portal that gives a description of the vital status and provides links from there to dive deeper into documentation on that. And the OMOP domain can be a useful tool when you're looking for research variables of interest.

The OMOP domain includes concepts using the OMOP concept table. So we can search the OMOP concept table for a concept that we're interested in. So for example, I'm interested in diabetes, and I'm limiting it to the idea of diabetes as a condition. And so you can see here it's returned a bunch of results for ICD codes related to diabetes. And I mentioned the VHA data portal. That is also another useful resource that provides links to documentation various VHA data that is not necessarily CDW data. And so you can use this as a tool to get at very particular data if that's what you're interested in.

All right, I'm going to talk about understanding patient identifiers a bit more. So I've been throwing out patient identifiers there so let's go into some more detail here about what patient identifiers you'll find in VHA data and how to use them. So you'll have noticed there are quite a few data sources, and there are quite a few different kinds of patient identifiers and so how do they all come together? So first, let's talk about the real social security number. So those familiar you'll know that generally a social security number has nine digits. It can have leading zeroes. If a social security number is not provided by a patient, then a pseudo social security number is assigned by the VistA system.

This will serve as a placeholder until the real SSN is obtained. So the pseudo SSN is a nine-digit number generated by the VistA system followed by the character P. The pseudo SSN that is generated is not unique and the same pseudo SSN may be assigned to different patients. Sometimes, SSN's may contain other formats, for example a 10-digit number, if the SSN was entered as such into VistA. Then we have our scrambled social security number. So this may be used as a replacement for the social security number in many VHA data stores for privacy and security reasons.

This is generated by CDW as well as other VHA systems using the same algorithm. It can be used to identify patients in some non-CDW domains, including MedSAS, DSS, Vital Status, and HERC. The scrambled social will be the same for the patient that has a pseudo social security number as well as the patient that does not have a pseudo social security number if the first nine digits of the patient's social security number is the name. The patients with the pseudo SSN will have a P following the first nine digits of their patient social, but there is no P at the end of the scrambled social.

Therefore, the scrambled social may not be unique for a small number of patients. We have the Sta3n or station so this is a three-digit number that represents the VistA system or station where the data was entered. So when using raw data that had not had any processing and using any IEN, you must use the station as well to ensure uniqueness, and I'll show some examples of these in the next few slides. So for example, you need to join on the patient IEN and the station.

So the patient IEN is found in both the Patient and SPatient tables. This is the internal entry number in the VistA system for a patient. It is unique to a patient in only one VistA system but not across all VistA systems. So again, as I mentioned, when using raw data that has not had any processing, and if you're joining on the patient IEN you must use this station as well. So for example, we have our patient John Smith. He is assigned a patient IEN of 100 as station number 438. We have our patient Beth Adams.

She is assigned a patient IEN of 100 as well but at station 660. So if we join just on patient IEN, you will match John Smith to Beth Adams, and we know they are not the same person. So we also need to match on the station. And we have our PatientSID, so again the SID stands for station identifier. So for the PatientSID, this is a patient identifier for CDW production. It is also in TIU and some other processed raw data. This is the primary key for the Patient and SPatient tables. This is generated by CDW for each unique combination of patient in VistA system or station.

So a patient may have more than one PatientSID if they have information in more than one VistA system. So our example here we have John Smith who is patient IEN number 100 at station number 438. So he is assigned a PatientSID of 12345 in this station. The same John Smith is assigned a patient IEN of 250 in station 660, so he is assigned a PatientSID of 8888 at station 660. So now, John Smith has two different PatientSID's. So that's why we have the patient ICN. The patient ICN is found in both the Patient and SPatient tables. It is used at the patient identifier for the CAN score domain. It is created by the Master Veteran Index or MVI. It is not reused and it unique across all patients and all VHA systems, including VistA systems.

So a patient should have only one ICN. So if a study requests data from a domain that uses scrambled social as a primary identifier, but they do not want scrambled social access and the regulatory that comes with that, patient ICN can be substituted as an identifier instead. So we have kind of a summary here of different identifiers. We have Jonathan Smith. This patient has been to three different stations, so you can see there he was assigned three different PatientSID's. His social security number is the same, and he is assigned the one patient ICN. So as a word of caution, similarly-named identifiers are not always the same thing.

For example, the PatientSID from the patient domain is not the same thing as the PITPatientSID from the PIT domain so please check the primary and foreign key relationships before you match. And now I'll talk just a little bit about indexing in SQL. So what is a SQL index? It's an on-disk structure that is associated with a table or view that speeds the retrieval of rows from that table or view. So bottom line, indexing is used to make queries more efficient and faster. It contains keys built for one or more columns in the table or view.

An index in SQL can be clustered or non-clustered. So a clustered index will sort and store the data rows within the table or view where as a non-clustered index is a structure that is separate from the data rows, and it contains the non-clustered index key values and has a pointer to the data row that contains that key value. So indexes and constraints so indexes are automatically created when a primary key and unique constraints are used on table columns. A primary key constraint will automatically create a clustered index, and a unique constraint will automatically create a non-clustered index.

And then just a note on a clustered columnstore index, so something that is neat about a clustered columnstore index is it combines the query efficiency of an index, and it adds data compression. So this is recommended for use on large tables you may have like fact tables that that to require full table scans and have a lot of data rows. All right, as I mentioned, when these slides are made available, I've included a bunch of links in my references and sometimes on the slides themselves. So all of these various tools and domain documentations I've referenced. So you can dive into more detail through these. If you ever have questions about anything VINCI related, please email us at vinci@va.gov.

Rob: Thank you, Brian. We do have a number of questions queued up and we'll launch right in, but first let me say, attendees, if you have questions for Brian, please send them to the Q&A. It's difficult to manage the chat. The chat only goes straight to me so he can't see the Q&A, but I'm going to read them as they came in. First off, what about DICOM files for patient imaging? These are often used in deep learning and other AI research. Is that accessible to identify for patients at scale?

Brian Robinson: So I am not familiar with that data actually, so I apologize. I don't know.

Rob: Thank you, moving on. This person asks what is the cadence/schedule for updating the MVI?

Brian Robinson: That is a great question. I don't know that off the top of my head, but I believe I linked the MVI. I need to add a link to MVI, don’t I? So I don't know that off the top of my head. I apologize. I know that it's considered the authoritative source for things like the patient ICN and keeping track of Veteran demographics and things like that, but I do not know if there is any kind of update schedule.

Rob: Thank you. They can send an email to vinci@va.gov, correct? Is that the best email to send questions to?

Brian Robinson: Yes. So if you have any questions, send them there. It will be allocated to someone who is more of an expert on whatever the subject is and we can find out for you.

Rob: Thank you. This person asks please advise where to find meta data definitions and explanations for each domain.

Brian Robinson: So probably one of the best tools for that will be through the BISL tool. So if you go to the BISL sharepoint site in my references here, you can go to their Meta data link, and then they have a Meta data report tool that is a power BI report. And that will breakdown CDW for you. You can select your CDW domain. It will show you the various tables in that domain and you can click through. And it provides Meta information, including column descriptions. It includes primary and foreign key relations, things like that.

Rob: I have a couple of questions here that are probably the same one. One person asks how do I access the Dim tables within VINCI? Another asks where do you go to even see the Dim data? I think that's probably the same question.

Brian Robinson: Yeah. So the Dim tables, because they do not have patient identifiers, they are available through the CDW work database. So if you query cdwwork.dim.viewname, there's a bunch of them in there, and I referenced a bunch of names for them SQL server management studio. You can also go to CDW work and see the Dim views listed there. Any research project will have access to these, so if you're working for an ORD, you can query the Dim tables directly from CDW work. So you should've been granted access to all of those Dim tables.

Rob: Thank you. I apologize to the questioner. This one came in early, and I missed it so I'm going to let is skip the queue. This person asks will this information be comparable/valid outside of VINCI (LSV), etc.

Brian Robinson: So I'm not super familiar with the LSV, but my limited understanding of it is that the data will be similar. So I would say because this presentation was so general, you might find some use there, but I don't know off the top of my head if it's a direct one to one of kind of how the data shows up in the LSV. I've not used it personally, so I apologize. I'm not that familiar with it.

Rob: Thank you. This is a long one so bear with me. As a site earlier today, we were looking for BMI with CDW work, and it did not find it. If we were looking in the places referenced on slide 40 (Dimvitatype.vitalsigns) then it seems like we are safe assuming that BMI is not available elsewhere in CDW work (despite the fact that BMI is listed on the CPRS coversheet along with all of the other vital signs)?

Brian Robinson: I can't speak to the CPRS site, but I can speak to, yes, you are not going to find BMI directly in CDW, but you're on the right track looking in the vitals domain. So in the vital measurements you can find the height measurements for patients. You can find the weight measurements for patients, and then you can use the BMI equation to calculate BMI. So you can get at BMI kind of indirectly that way by calculating it yourself.

Rob: Thank you. We have a couple questions about where the slide link is. I've posted it multiple times to the chat, and I can do so in a moment while Brian is answering this next question. Will this be helpful at all in trying to find patient-specific information say about smoking history or chemical exposure or when the patient was in the service whether they were in Vietnam?

Brian Robinson: Yeah, so a couple different domains you're going to want to look into are first the health factors domain is a great starting off point for looking for smoking information. There is quite a bit of literature on this topic of kind of identifying smoking within the VA so you can also see what other teams have done. Another useful resource -- and I've linked it in the references here on one of these slides -- would be decipher wiki. Let's see, right here, the VA Phenomics Library. So this link you can go there and see what others have done to kind of get at smoking within CDW data.

So if you search for smoking as a phenotype, you'll find what they’ve done there. I believe the Covid shared data resource has also dug into smoking, but I could be wrong. So that's another way to look, but if you're just wanting to dive into yourself, health factors is where you'd want to start for smoking status. And then for service-related information, that's going to fall in the patient domain so the Patient and SPatient tables that include information on service period and other service-related information would be there.

Rob: Please remind me what Dim table/view is.

Brian Robinson: So the data within VA is accessed through SQL, and so in SQL a view is -- sometimes we use the term view interchangeable with table. There is a distinction between them, but generally they behave the same. The view is just pointing at an actual underlying table. Most of the data provided in the VINCI workspace comes through views, and so when we say view or we say table, we use the term interchangeably. And at a most basic level, you can do so and they behave the same.

So the idea is that there are a bunch of views or tables provided that are referenced as Dim tables because they don’t have any patient data in them so they're dimension or Dim tables, and then we have the rest of the data is in the tables, which are known as fact tables because they do contain patient information. So the Dim tables you'll find a bunch of different data, but none of it is patient data. It's things like ICD codes or medications or data that specifies those kinds of things without actually getting into patient data.

Then you can use those to match to the actual fact tables where the patient data is to kind of get a picture of whatever you're interested in. So say you're interested in ICD diagnoses. You can start in the Dim table to look at the ICD diagnosis codes and descriptions. Then you can join to the fact table to determine which patients had the diagnosis codes that you're interested in.

Rob: Is it possible to get a demo of using the table value function TVF\_TIU\_fulltextsearch to search TIU notes?

Brian Robinson: I don't know off the top of my head if that's been presented in another seminar or any kind of resource there. If you would like kind of some one-on-one help with that, please email us at vinci@va.gov and someone should be able to kind of walk you through how that's used.

Rob: Thanks Brian. Does MedSAS refer to CMS data?

Brian Robinson: Yes.

Rob: How does research staff gain access to CDW?

Brian Robinson: Good question. If you start from the VINCI central website, we have some links there kind of directing you of how to start the process. So we have a bunch of useful potential tools for you before you even request access. So we have a bunch of preparatory to research tools like you can request a feasibility or a data needs assessment before you even apply for access to kind of narrow you in on what exactly do I need to get access to? Which data should I be considering?

And then you'll go through the DART tool or data access request tool, and that's available through VINCI central. And through there, you'll go through the regulatory process of applying for access, providing any required regulatory information, and once that is approved you'll be granted access to the CDW data and other data you’ve requested.

Rob: This person asks how do you search free text areas in the health factory domain?

Brian Robinson: So you'll start with the Dim table. So there is a Dim table in the health factor domain, and let's see if I listed it straight up or not. There we go dim.healthfactortype. So start your search there, and there is a bunch of free texts that list the health factor type that is assigned to the various health factors.

So you'll determine which health factor types you're interested there, and then once you’ve identified those, you'll use the -- I don't know what it is exactly off the top of my head -- it'll be like the health factor type SID or something of that nature, which will be a primary foreign key joined into the health factor table, the actual fact table to pull the patients who have that health factor that you're interested in.

Rob: Do you have any recommendations on resources on how to figure out the arc (vera) data structure, which might not be a data source within CDW or please correct if I can find it within CDW?

Brian Robinson: I apologize. I'm not familiar with that one so I do not know.

Rob: Is there a good contact or source for PIT data, particularly linking patient ID's, patient account numbers to the outpatient and other CDW files?

Brian Robinson: So I know BISL has some documentation on the PIT or PIT domain so I would start there, and I've linked it kind of at my bottom in my references. So start with the BIS, look into their Meta data, look into their domain release information. That's where I am familiar and know of various PIT documentation to hopefully kind of point you in the right direction regarding how to utilize the PIT domain.

Rob: Just a few more left. Does the PSSG data contain census-blocked data?

Brian Robinson: That is a great question. I'm not sure off the top of my head. I know I've gotten asked census-blocked data before. I don’t remember where I did it off the top of my head, and so I apologize. Any of these questions that I haven't been able to answer because I'm only one person and I'm hardly the expert on everything please email us vinci@va.gov and ask your question and we'll get back to you.

Rob: Thank you. We have two more. Where can we find JLV medications? Do I need to access to DaVINCI to do so?

Brian Robinson: So JLV I'm not familiar with that. I apologize.

Rob: That person can write in on classification on what JLV. We really don’t have too much time. This person asks do we have a forum of researchers who use CDW? Do you know the answer to that question?

Brian Robinson: That I do not know.

Rob: It looks like a couple may have come in just at the end. This person writes a comment just a comment on the vera arc question. Our health informatics at our VA do use SQL to pull vera data from the CDW. So it's in there somewhere. Another person asks please let attendees know OMOP has recently improved smoking data. If you have access and want help, contact omop@vinci@va.gov. I think that's really about it Brian and all the time that we really have so if you'd like to make closing comments now is your chance.

Brian Robinson: So as I mentioned at the beginning, I know this was very high level and very general. I'm hardly the expert on everything like I said, so if your question wasn’t answered and you want to learn more, please email us at vinci@va.gov. And people that would know more can reach out to you. We can help you. We're here to help. We want your questions, and we want you to get your questions answered so we'll do whatever we can to help so go ahead and shoot us an email again. It's just vinci@va.gov and that will be directed to the team or person that would best be able to assist you.

Rob: Great, thank you, Brian Robinson. That's vinci@va.gov, and attendees I want to close the webinar momentarily. A webpage with a few survey questions will pop up. Please spend a few moments and provide answers to those questions. We really do appreciate you taking the time. Thank you once again, Brian. With that, I'll just wish everyone a good day and go ahead and close.