The following is a listing of clinical research databases for which BDSC can provide support services. In addition, if you wish to do your own statistical analyses but do not have statistical software loaded on your computer, you can access a wide variety of software programs through Penn’s library. This software can be accessed at the [Penn Library](https://guides.library.upenn.edu/vlab)

The **IBD Immunology Initiative** (the I3 Study), is a comprehensive research study at the University of Pennsylvania Health System created to better understand the connection between IBD and the immune system. It is a joint study initiative between the Division of Gastroenterology, the Institute for Immunology at the University of Pennsylvania, and the CMSDLD. The I3 Study aims to enroll every patient with Crohn's disease, ulcerative colitis, and indeterminate colitis seeking care at Penn in order to create a biobank of tissue and blood samples as well as clinical information. To date, more than 1500 patients have been enrolled, consented to allow for chart reviews and biosample collection, and their data are stored in a REDCap data base, including key variables such as medication and surgical history and disease phenotype. Among these, 788 have banked tissue samples (i.e. mucosal biopsies, plasma, lamina propria mononuclear cells (LPMCs), peripheral blood mononuclear cells (PBMCs), and organoids).

* 1. Contact person for access to data or biosamples: Meenakshi Bewtra
  2. Cost to access data: Negotiable depending on types of services requested and funding of project
  3. Cost to access biosamples: Negotiable depending on type of biosample requested (blood vs tissue) and processing requested
  4. IRB requirements: I3 has an existing IRB protocol #814428. If Dr. Bewtra or Dr. Wherry are PIs on the project, an amendment to the I3 is possible. If the proposed project is a separate project, a separate IRB must be submitted and can reference the I3.  Our team can assist you with this submission.
  5. Years of available data: 2011 - Present
  6. Special considerations: Please see website for additional details and contact us if you have any questions

1. **SPARC IBD** is a prospective cohort study of adult patients with IBD conducted at 17 institutions (and growing) across the USA. With more than 5000 patients enrolled, of whom more than 500 were recruited from the University of Pennsylvania, SPARC-IBD is one of the most unique IBD registries in the world. Each of the participating sites follows a standardized protocol for collection of patient-reported outcomes and biosamples, including PBMCS, plasma, stool, and mucosal biopsies (formalin-fixed paraffin-embedded (FFPE), RNA later, liquid nitrogen). Clinical data include demographic data, symptoms, endoscopic assessments, laboratory data, hospitalizations, prescription medications, diagnoses, etc. Two particularly unique features of the SPARC IBD cohort are that the participating patients’ electronic health record data are shared within a common data warehouse and that more than 50% of the cohort members have their data linked to commercial insurance claims data.
   1. Contact person for access to data or biosamples: Jim Lewis
   2. Cost to access data: no cost for SPARC data. There is a cost for obtaining linked commercial claims data. Individual pricing depends on the scope of the project and is determined by the commercial data vendor.
   3. Cost to access biosamples: no cost
   4. IRB requirements: There is an existing IRB protocol that covers retrospective analyses of the SPARC data for questions related to IBD.
   5. Years of available data: 2017-2023 for prospective data. There are some patients with electronic health record data prior to enrollment.
   6. Special considerations:
      1. Access to biosamples requires a formal application to the IBD Plexus review board. The number of samples requested is carefully scrutinized to assure adequate power but to avoid excess use of limited resources. All data generated from the biosamples must be deposited back into the SPARC database after a limited period of exclusivity.
      2. A large number of patients samples have been used to generate microbiome sequencing (bacterial and fungal) and bulk RNAseq data that is available at no cost
      3. The University of Pennsylvania receives a fresh data cut of the SPARC data approximately twice per year.
      4. Data may not be shared with collaborators outside of the University of Pennsylvania without putting a Data Use Agreement in place with the Crohn’s & Colitis Foundation
2. **University of Pennsylvania Health System’s (UPHS)** **Epic electronic health record** and other electronic data included in Penn Data Store. UPHS was an early adopter of the Epic electronic health record. There are nearly two decades of data stored within the UPHS Epic Clarity database. Other electronic databases (e.g. endoscopy, radiology and pathology) are linked to the Epic system through Penn Data Store. These data are most efficiently accessed through CDOTS who can generate queries of the data and provide access to these data for research.
   1. Contact person for access to data or biosamples: Marina Serper
   2. Cost to access data: no cost
   3. Cost to access biosamples: not applicable
   4. IRB requirements: There is an existing IRB protocol that covers retrospective analyses of UPHS for questions related to gastrointestinal and liver diseases (EAGLES protocol). Investigators need to be added to the protocol to use the data.
   5. Years of available data: beginning of Epic at Penn to real time
   6. Special considerations:
      1. Data may not be shared with collaborators outside of the University of Pennsylvania without putting a Data Use Agreement.
      2. **Epic Clarity Registries**: Within Epic, there are a number of unique registries that have been developed. For example, the **PenNSAM** **Medical** **Nutrition Registry** in EPIC contains detailed nutrition information on hospitalized paitents. Due to the foresight of our dietitian service, all of the clinical dietitian’s assessments for malnutrition among inpatients at Hospital of University of Pennsylvania have been recorded in an electronic database using defined vocabulary. These data have been merged with the Epic record to create the **PenNSAM** Medical Nutrition Registry.
      3. **Epic IBD Smartform**: The Epic IBD Smartform is a phenotyping tool developed by by Penn investigators to allow real time phenotyping and tracking of symptoms among patients with IBD seen in UPHS clinics. This tool has now been made freely available to all Epic users. An add on to this tool is a pre-visit symptom survey tool that can be used to systematically collect patients symptoms via internet surveys or using tablet computers in waiting rooms. This tool is utilized for all follow-up visits for patients with IBD seen at UPHS. The tool improves both the quality and speed of data collection by reducing the need for manual chart reviews or use of tools such as natural language processing.1 This tool has been fundamental to the implementation of the I3 and SPARC IBD cohort studies described above. Continued refinements to the tool are managed by Dr. Lewis in collaboration with Penn’s Epic Support Team

1. **Penn Medicine Biobank (PMBB):** The PMBB supports researchers by providing centralized access to a large number of annotated blood and tissue samples. To date over 60,000 participants have enrolled in the PMBB. PMBB participants range in age from 18 to over 100 years old. Racial and ethnic diversity mirrors the patient population at the University of Pennsylvania Health System. Consent for the PMBB is obtained with a consistent informed consent document, which includes permission to re-contact participants for future research opportunities and to collect additional follow-up and outcomes data. Blood samples are banked as whole blood, plasma, serum, buffy coat, and DNA for future studies following stringent standard operating procedures. Clinical data are obtained through multiple sources, such as, a patient questionnaire filled out at the time of the recruitment, electronic medical records for abstraction by trained study personnel and the Penn Data Store.
   1. Contact person for access to data or biosamples: JoEllen Weaver (or email [biobank@upenn.edu](mailto:biobank@upenn.edu))
   2. Cost to access data: Variable by project
   3. Cost to access biosamples: Variable by project
   4. IRB requirements: An IRB approval is required to access biosamples
   5. Years of available data: Enrollment in PMBB starting in 2013. Biosamples in PMBB can be linked to clinical data from UPHS as described in the section on UPHS.
   6. Special considerations:
      1. To gain access, see <https://pmbb.med.upenn.edu/data-access/gaining-access.php>
      2. A dashboard can be accessed to gauge the size of available populations and samples: <https://ritchielab.org/pmbb-apps/dashboard/>
2. **Insurance data from Medicare**: The University of Pennsylvania houses portions of data from Medicare (government insurance for elderly and those with specific medical conditions) for use by clinical investigators. Medicare data come in multiple different files, such as those for inpatient encounters, outpatient encounters, medications, durable medical goods, etc.
   1. Contact person for access to data: Jibby Kurichi
   2. Cost to access data: The cost to access data varies based on how the user chooses to access the data. Some Medicare data are stored on servers managed by Leonard Davis Institute through their HSRDC service. For these data, there is a cost for server space depending on funding source, the PIs level, and project size. In addition, there is a $2000 fee for a data reuse agreement which must be paid to CMS. If data beyond those stored at Penn’s HSRDC are needed, there will be a cost for creating those data files. A list of data available on the HSRDC can be viewed at: <https://www.dropbox.com/scl/fi/h7bhfhw9gzlsf6ygjdre2/Current-data-on-the-HSRDC.xlsx?rlkey=jtqlz70z99pzi6pk5ii9p6ch5&dl=0through>  The CMS website can help you estimate that cost of obtaining additional files and working with ResDAC can get you a more exact estimate.

The alternative way to access data are through a “seat license” to VPN into data stored on the CMS research servers at the Virtual Research Data Center (VRDC). This is the preferred method to access the data as it avoids the need for individual data use agreements and assuring appropriate data custody arrangements at Penn. Information is available at <https://resdac.org/cms-virtual-research-data-center-vrdc> and are summarized below.

**Seat Access -** Researchers who access data in the secure VRDC environment will be charged a standard access fee per user or “seat.” This fee covers CMS onboarding, seat license, training, output review, and administrative costs. The seat access fee is charged on an annual basis; each seat must be renewed every year in order for the user or “seat holder” to continue working on the study.

**Project Fee -** The VRDC project fee is a one-time standard fee that covers the cost of extracting data needed. There is no charge to add additional years of data for an existing cohort. However, any changes in the cohort that result in re-extracting data will incur a fee. Existing VRDC seat holders may add projects to their user workspace for an additional project fee.

**Space Cost -** The first 500 GB of space are included in the project fee. However, researchers may need to pay for additional space in the VRDC depending on the size of their data request. Space is needed for raw data, analytic files, and output. Additional space can be purchased in 500 GB blocks. The cost for continued additional space will be charged during the seat renewal period, if applicable.

Obtaining a formal cost estimate is the best way to be sure of all associated VRDC fees, including whether or not extra space is required. See the FAQs for details at <https://resdac.org/sites/datadocumentation.resdac.org/files/2022-02/CMS%20Fee%20List%20for%20CCW%20VRDC%20Cloud%20Environment.pdf>.

* 1. Cost to access biosamples: not applicable
  2. IRB requirements: An IRB approved protocol will be required to use CMS data
  3. Years of available data: The HSRDC Medicare data varies by data file and can be seen at: <https://www.dropbox.com/scl/fi/h7bhfhw9gzlsf6ygjdre2/Current-data-on-the-HSRDC.xlsx?rlkey=jtqlz70z99pzi6pk5ii9p6ch5&dl=0through>. Using a seat license, it is possible to update your files when new data become available at no cost.
  4. Special considerations:
     1. Investigators are encouraged to seek a consult from ResDAC at University of Minnesota for help with selecting data files, implementing a DUA, etc. There is no chare for the ResDAC service.
     2. Investigators are prohibited from reporting cell counts <11
     3. Some patients have both Medicare and Medicaid benefits. As such, some investigators use both data sets to fully capture this dual eligible population.
     4. Medicare and Medicaid data can be linked to National Death Index data
     5. Medicare and Medicaid data can be linked to other datasets with appropriate protections in place. For example, Medicare data have been linked to the national organ transplant registry data.
     6. Medicare data are linked to the SEER cancer registry and are available for researchers

1. I**nsurance data from Medicaid:** The University of Pennsylvania does not house data from Medicaid (government insurance for low income and special-needs individuals), but some investigators have DUA agreements for Medicaid data that could be leveraged for a data reuse agreement. Alternatively, Medicaid data can be accessed via the VDRC as described above for Medicare
   1. Contact person: would need to find a DUA holder at Penn or go through the VDRC at CMS.
   2. Cost: same as for Medicare as described above
   3. IRB requirements: IRB approved protocol is required to access data
   4. Special considerations:
      1. Investigators are encouraged to seek a consult from ResDAC at University of Minnesota for help with selecting data files, implementing a DUA, etc. There is no charge for the ResDAC service.
      2. Investigators are prohibited from reporting cell counts <11
      3. Some patients have both Medicare and Medicaid benefits. As such, some investigators use both data sets to fully capture this dual eligible population.
      4. Medicare and Medicaid data can be linked to National Death Index data
      5. Medicare and Medicaid data can be linked to other datasets with appropriate protections in place. For example, Medicare data have been linked to the national organ transplant registry data
2. **Insurance data from commercial insurers**: The University of Pennsylvania houses data from from OptumInsight Clinformatics (Optum), a patient-level medical claims database consisting of the inpatient, outpatient, pharmacy, procedure, and laboratory claims of more than 60 million unique patients from a large national health insurer from 2000 to 2021. Optum also contains claims from patients co-insured with Medicaid from 2000 – 2010 and Medicare from 2000 – 2005.
   1. Contact person for access to data or biosamples: Jibby Kurichi
   2. Cost to access data: Users incur a cost for server space through LDI that is based on the percent effort that programmers will be working on the project. If the research is funded by a government grant and the direct costs are greater than $200,000, there is a one time charge of $40,000. If the direct costs of the government grant is less than $200,000, the fee is the lesser of $15,000 or 20% of the direct costs of the grant.
   3. Cost to access biosamples: Not applicable
   4. IRB requirements: Because the data are completely deidentified, research using Optum data is considered exempt from requiring IRB approval.
   5. Years of available data: 2000-2021
   6. Special considerations:
      1. Optum data may not be linked with other data
      2. Funding for Optum research may not come from industry sponsors.
      3. All abstracts and manuscripts must be approved by Optum prior to submission to assure accurate description of the data source
      4. All grant applications must be approved by Optum prior to submission o assure accurate description of the data source
      5. The database includes all diagnoses and procedures that are recorded using ICD, CPT and HCPCS codes. Provider specialty can be identified. Geography is limited to regions of the United States. Data on race and other social determinants of health are estimated based on statistical algorithms.
      6. OPTUM data at Penn do not include data on death.
      7. The Biomedical Analysis Center (BAC) has extensive experience working with Optum data.
      8. Access to Optum data will eventually switch to a VPN model where users will log into a server managed by Optum. The timing and cost of this has not yet been determined

1. **IQVIA Medical Research Database (IMRD):** Previously known as The HealthImprovement Network (THIN), IMRD is a collection of medical records from primary care practices within the United Kingdom. The population included represents approximately 6% of the UK population. IMRD data include demographics, diagnoses (coded with READ codes), medications and SDOH-related data. Because most prescriptions in the UK are written by primary care providers, IMRD is a valuable tool for pharmcoepidemiology research. The dataset that contains clinical patient data recorded from UK practices up to12th January 2021. The dataset consists 19,508,644 patients with 1,479,318,232 encounter records from 832 practices.
   1. Contact person for access to data or biosamples: Danielle Mowery
   2. Cost to access data: covered under Penn’s license
   3. Cost to access biosamples: not applicable
   4. IRB requirements: IRB approval required. IQVIA Scientific Review Committee approval is required.
   5. Years of available data: Pre 2000 through 2021
   6. Special considerations:
      1. Some specialty medications, such as biologics for IBD, are incompletely recorded because these are more likely to be prescribed by specialists
      2. No further updates will be obtained by UPenn for THIN data
      3. See <https://ibi.med.upenn.edu/imrd/> for forms and additional information.
2. **Veterans Affairs Health System Data:** The VA health system contains data for over 9 million Veterans from 128 Veterans Affairs Medical Centers across the US and US territories. Data available from the VA are unique in that they are extremely granular—all patient-related interactions are captured in the dataset, which includes (among other data): demographics, longitudinal body mass index, smoking, and alcohol use data (via AUDIT-C), administrative coding data, hospitalization data, outpatient visits data, consult and referral data, pathology and imaging data, and detailed inpatient/outpatient pharmacy data. The VA health system cohort has a high burden of psychosocial and cardiovascular comorbidities, and is an excellent resource for natural history studies.
   1. Contact person for access to data:

Maria Souden, MSI, PhD  
(708) 202-2476  
email: [Maria.Souden@va.gov](mailto:Maria.Souden@va.gov)   
web site: <http://www.virec.research.va.gov/>

* 1. Cost to access data: none
  2. Cost to access biosamples: not applicable
  3. IRB requirements: must use the VA IRB
  4. Years of available data: 1999 to present
  5. Special considerations: A potential limitation is that the VA is male predominant which may impact external validity

1. **Vizient Clinical Data Base:** The Vizient Clinical Data Base (VCDB) is a collection hospital-level data on patient outcomes — such as mortality, length of stay, complication and readmission rates, and hospital-acquired conditions — that enable hospitals to benchmark against peers. It is designed for quality improvement but is also used for empiric health services research. The VCDB includes discharge diagnoses and billing data from approximately 450 hospitals across 41 states. Hospitals include both academic and community hospitals. Individual hospitals can be identified, but all other data are completely deidentified.
   1. Contact person for access to data: Craig Kean (Craig.Kean@pennmedicine.upenn.edu)
   2. Cost to access data: No cost
   3. Cost to access biosamples: not applicable
   4. IRB requirements: IRB approval may be required depending on the study design and the data that will be used.
   5. Years of available data: October 2019 through June 2023 for prospective data. There are some patients with electronic health record data prior to enrollment.
   6. Special considerations:
      1. VCDB does contain data on medication use
      2. VCDB has their own risk adjustment tool with relatively high accuracy, but it differs from tools used by CMS.
      3. Users of the Vizient data are prohibited from identifying individual institutions in publications.
2. **TriNet-X**
   1. Contact person for access to data: Laura Fluharty (lauraee@upenn.edu) and Yulia Borovskiy (Yuliya.Borovskiy@pennmedicine.upenn.edu)
   2. Cost to access data: none to PI as covered by Penn’s contract
   3. Cost to access biosamples: not applicable
   4. IRB requirements: Must provide IRB approval letter or non-human subjects determination
   5. Years of available data: 2005 to present
   6. Special considerations:  Datasets are deidentified or Penn only data, includes dates. Dataset requests need to go through REDCap tracker to the DNA team and then will be available for download or brokered. Does not include research opt outs, HIV data, behavioral health data.