### Medical Informatics and AI (MIAI) Core

# [Short Version]

The Department of Biomedical Informatics Medical Informatics & AI (MIAI) Core

The Medical Informatics & AI (MIAI) Core provides interrelated services to assist investigators across Emory University:

**SpringBoard** – A consulting service for investigators seeking insight and input on acquiring, storing, and using data in their investigations. SpringBoard provides insight on self-management and core service recommendations to ensure that investigators are appropriately resourced to complete their study with adequate support and statistical power. SpringBoard consultations provide the maximum benefits during the conceptualization phase of a study but are also available throughout an investigator’s work whenever data issues arise.

**CohortCount** – The CohortCount service assesses patient population size at Emory and affiliated facilities matching the proposed study design of the investigators. This service is designed for an investigator who may have an idea for a study but do not have the ability to determine whether there is a sufficient patient population to conduct the study. The Core’s CohortCount service provides guided, consultative acquisition of a total number of candidate patients for such a study based on multiple variables such as those available in the EMR or other clinical databases. Facilities accessible through this service include Emory Healthcare, Children’s Healthcare of Atlanta, Grady Memorial Hospital, and the VA.

**DataDig** – The DataDig service is designed to assist investigators with data extraction from clinical databases in a collaborative effort with the information technology departments at each health system and the clinical domain experts. Extracting data from EMRs and other clinical databases is a complex and arduous task a often requiring extensive amounts of time, technical credentials, familiarity with medical data, and specific knowledge of research activities. This service addresses these challenges through collaborations with clinical investigators at each facility that are charged with being liaisons to such data and BMI staff dedicated to extracting, collating (and deidentifying data if appropriate) as required. There is often the need to merge data from multiple sources prior to analysis. This can be the simple integration of data from two separate EMRs or more complex integration involving external non-medical data sources. The data integration component of the DataDig service provides the necessary insight to generate comprehensive datasets from multiple entities that are ready for cleaning and analysis. It also provides services such as normalization, ontology encoding, deidentification, and advice on encoding, representation and informatics pipelines, including cloud services. Once a dataset has been fully extracted and created there are frequent concerns regarding elements that are erroneous, improperly converted or difficult to interpret. This service may also include deidentification. The data cleaning component of the DataDig service provides necessary insight and technical expertise in cleaning the data using the expertise of a combination of clinicians, research investigators, and staff.

**DataGrab** – Sometimes investigators need data from devices or resources beyond standard medical databases, such as from wearables, mobile phones, social media, enduring sensors etc. The DataGrab service is designed to address these situations. Using Core resources, BMI staff and investigators are able to extract and develop comprehensive datasets from diverse data sources.

**RealTimeData** – Although many investigations are amenable to retrospective data or data archived in various sources, there are some situations in which data must be aggregated and presented in near real-time. This includes enrollment in clinical trials or testing of new technologies/informatics methods. The RealTimeData service provides the necessary staff to develop, implement, and support such systems for research and development purposes.

**DataHack** – BMI faculty and staff represent a valuable resource of state-of-the-art knowledge in the analysis of medical data, ranging from the application of signal processing to deep learning. This service provides advice on which experts to work with and potential techniques to employ on data.

**App Development/Backend** – Development services can assist researchers by providing full stack (mobile application + web dashboard + backend) development services. Our team of software developers and project managers can also provide related services such as initial consultation, prototyping/conceptualizing and project management during the duration of our development.

**Synthetic Data** – It may be impossible to use only data extracted from eMR and related data sources due to security and privacy constraints, particularly in small patient populations leading to high re-identification risks, or due to constraints in the study design and regulatory approval. The Synthetic Data Service creates artificially generated datasets based on specified data distributions. Such synthetic datasets may be used independently or as supplement to real datasets, and are well suited for development and testing of computational methods including AI/ML modeling, and as an alternative to de-identification for data sharing.

**Visual Analytics** – Once a dataset has been fully extracted and created there are frequent concerns regarding elements that are erroneous, improperly converted or difficult to interpret. This service may also include deidentification. The Core’s DataDetox service provides necessary insight and technical expertise in cleaning the data using a combination of clinicians, research investigators, and staff.

**ML Algorithm Development** – The Core, in conjunction with BMI faculty and staff, have extensive AI and ML expertise. The ML Algorithm Development service provides customized algorithm development capabilities, as well as deployment for EPIC.

**Infrastructure Support/Maintenance** – The Infrastructure Support and Maintenance service provides hardware and software management support, in coordination of BMI IT support and Emory OIT. The core provides services ranging from hardware and software recommendation for application deployment and their configuration, hardware and software licensing cost estimates, deployment, and maintenance. We have expertise in working with physical as well as virtualized hardware such as VM and containers, data storage, OS and application deployment and maintenance, and CPU and GPU high performance computing, both on-prem or in-cloud. The core in addition provides data upload service, at regular intervals or on- demand, for projects with data sharing requirements.

**Project Management** – Project management services (for large MIAI requests) specialize in planning and coordinating of project activities, including core service activities, according to project-specific requirements and constraints, tasks and timelines. Project management services are available for the project lifecycle from conceptualization to completion, in part or in its entirety.

**GrantGen** – This service provides researchers with tailored text for grant proposals and project reports on Biomedical Informatics pipelines, including data acquisition, processing, infrastructure, and relevant statistical summaries.

# [Long Version]

The Department of Biomedical Informatics Medical Informatics and AI (MIAI) Core
<https://www.cores.emory.edu/miai/services/index.html>

The Medical Informatics & AI (MIAI) Core at Emory University provides an integrated suite of services to support biomedical, clinical, and translational research across the university and its affiliated healthcare institutions. Drawing on deep expertise in biomedical informatics, data science, and AI, the Core assists investigators with acquiring, managing, analyzing, and deploying data from a wide range of sources—including EMRs, wearables, imaging, unstructured clinical text, and real-time streaming data.

With support from dedicated engineers, core faculty, and domain experts in the Department of Biomedical Informatics (DBMI), the MIAI Core delivers end-to-end solutions throughout the research lifecycle—from study design to data acquisition, model development, and application deployment.

### Core Services

SpringBoard – A consulting service that helps investigators plan how to acquire, store, and use data for research. SpringBoard provides tailored guidance on data strategy, self-management, regulatory pathways, and statistical power, ensuring each project is set up for success. Faculty and staff can receive up to 2 hours of complimentary consultation. SpringBoard also helps connect investigators with the appropriate BMI experts based on study needs.

CohortCount – Assists investigators in evaluating the feasibility of clinical studies by estimating the size of relevant patient populations. Using EMR data from Emory Healthcare, Children’s Healthcare of Atlanta, Grady Memorial Hospital, and the VA, this service provides guided estimates based on inclusion/exclusion criteria and clinical variables of interest.

DataDig – Facilitates the extraction of structured and unstructured data from clinical systems. DataDig includes services for retrospective data extraction, multimodal data integration, normalization, ontology encoding, de-identification, and data cleaning. It supports a wide range of data formats, including clinical notes, physiological waveforms, imaging, lab results, and more.

DataGrab – Enables access to data from sources beyond traditional medical systems, including wearable devices, mobile phones, sensors, social media, and public databases. DataGrab is designed for studies investigating real-world data, digital health, and social determinants of health.

RealTimeData – Provides infrastructure and technical support for studies requiring near real-time data access, such as clinical trial recruitment, remote monitoring, or live decision support. This service enables continuous data feeds and dynamic model retraining pipelines.

Synthetic Data and Digital Twins – Offers the generation of synthetic datasets that preserve the statistical characteristics of real-world data. These synthetic datasets can supplement real data, support algorithm development, or serve as privacy-protected alternatives for sharing and testing. Digital twin generation enables simulated, patient-specific modeling for research or clinical innovation.

DataHack – Connects researchers with BMI experts for short-term feasibility assessments or advice on selecting and applying appropriate analytical techniques. Services range from classical statistics to signal processing, machine learning, and deep learning.

Visual Analytics – Supports the development of interactive dashboards and data visualizations for real-time monitoring, exploration, and dissemination. This includes identifying anomalies, addressing data quality issues, and facilitating interpretability of complex datasets.

Machine Learning Algorithm Development – Offers customized ML and deep learning model development, including data preparation, model training, validation, testing, and deployment. This service includes support for integration with platforms such as EPIC or bespoke research applications.

Natural Language Processing & Large Language Models (NLP & LLMs) – This service provides a comprehensive set of tools and expertise for working with free-text clinical data and biomedical language tasks. The Core supports fine-tuning and customization of open-source large language models (LLMs) for research purposes, including domain-adaptive and topic-specific training.

Services include high-precision de-identification of clinical notes, named entity recognition (NER), and information extraction using both rule-based and machine learning-based approaches. The team also offers support in building and expanding clinical lexicons, developing regular expressions for complex text patterns, and performing fuzzy matching for inexact text searches.

Investigators can leverage NLP tools to discover patient cohorts from unstructured clinical notes, especially when traditional coding systems (e.g., ICD) fall short. Prompt engineering strategies—including chain-of-thought and retrieval-augmented generation (RAG)—are also available to improve LLM outputs for specialized tasks.

Additional services include supervised classification of clinical text, creation of end-to-end NLP pipelines, and feasibility consultations to assess project readiness for NLP/LLM integration. All NLP and LLM services are delivered within HIPAA-compliant infrastructure to ensure data security and regulatory compliance.

Application Development & Backend – Full-stack development of mobile and web-based research applications, including user interfaces, secure backend systems, and dashboards. The team supports prototyping, technical scoping, and project management throughout development.

Infrastructure Support and Maintenance – Provides researchers with recommendations, setup, and ongoing support for cloud-based or on-premise infrastructure. Services include assistance with hardware and software selection, VM/containerized environments, high-performance computing (CPU/GPU), and secure data upload services.

Project Management – Offers expert project coordination services for studies involving multiple MIAI components. Support includes timeline development, resource allocation, team coordination, and deliverable tracking. Available for all phases of a project—from grant submission to execution and closeout.

GrantGen – Supports investigators in crafting informatics-related sections of grant proposals, IRB protocols, and progress reports. Text is tailored to the specific research aims and includes detailed descriptions of data pipelines, analysis methods, and infrastructure requirements.

Software and Model Deployment – Ensures that ML, NLP, and LLM models are securely and efficiently deployed into research or operational environments. Services include containerization, hosting on the BMI high-performance computing (HPC) infrastructure, and integration with external systems. Planned services also include model validation support aligned with FDA Medical Device Development Tools (MDDT) guidelines.

Data Management – This service provides comprehensive support for organizing, storing, and maintaining research data throughout its lifecycle. The Core assists investigators in formatting data to be AI/ML-ready, ensuring that datasets are structured and standardized for analysis and modeling.

Support is available for both secure on-premises and cloud-based storage solutions, tailored to project-specific needs and compliance requirements. The Core offers guidance on data governance, including alignment with IRB protocols, data use agreements, and institutional policies.

Investigators can also receive assistance in preparing Data Management and Sharing Plans that meet NIH guidelines. Additional services include long-term data lifecycle planning, archiving strategies, and support for submitting datasets to public repositories as required by funding agencies or journals.

To learn more or request services, please visit: <https://www.cores.emory.edu/miai/services/index.html>For direct inquiries:ci.core@dbmi.emory.edu