

**Course Name:** MS Health Informatics and Data Science (HIDS)

**Course Number:** HIDS 7950: Mandatory Capstone, 4 credits

**Contact persons:**

*Course Director*: Adil Alaoui, MS, MBA

[alaoui@georgetown.edu](mailto:alaoui@georgetown.edu), (202) 687-9108

2115 Wisconsin Ave, NW, Suite 200, Washington, DC 20007

*Program Director*: Yuriy Gusev, PhD

[yg63@georgetown.edu](mailto:yg63@georgetown.edu)

2115 Wisconsin Ave, NW, Suite 300, Washington, DC 20007

**Course Dates:** Summer Semester - May 19, 2025 - August 5, 2025

| **Student Information** |
| --- |

**Full Name of Student:** Austin Zach Cherian

**Address**: 2610 Tunlaw Rd NW, Apt #301, Washington, D.C. 20007

**Telephone**: (414)-573-0047 **Email:** azc10@georgetown.edu

**Emergency Contact:**

**Name:** Reena George **Relationship:** Mother

**Telephone:** (414)-550-6602

| **Capstone Mentor Information** |
| --- |

**Organization name: Georgetown University –**

**Organization Co-Mentor Name and Title**: Neil Shah

**Email:** shahn6@mskcc.org

**HIDS Co-Mentor Name and Title:** Yili Zhang, PhD

**Email:** yz932@georgetown.edu

| **Capstone Project Information** |
| --- |

**Capstone Project Description:** This project focuses on leveraging the power of Large Language Models (LLMs) to automate the extraction of detailed information about immune-related adverse events (irAEs) from unstructured clinical notes. This addresses the significant challenge and effort involved in manually curating this data from EHRs, which is crucial for research

and clinical decision support in immunotherapy. The project directly aligns with Aim 1

described in the research strategy document, particularly the plan to utilize LLMs for clinical Named Entity Recognition (NER).

**Objective:** To develop and evaluate a Large Language Model (LLM)-based pipeline that automates the extraction of immune-related adverse events (irAEs) and other clinically relevant features (e.g., social determinants of health, first-line therapies, treatment timing) from unstructured clinical notes in the IO Registry. This will reduce the need for manual curation and support downstream predictive modeling to improve outcomes in immunotherapy.

**Outcomes:**

* An NLP pipeline using ClinicalBERT or a similar fine-tuned LLM for Named Entity Recognition (NER) of adverse events and treatment details.
* Accurate extraction of irAEs from clinical notes, benchmarked against clinician-labeled ground truth.
* Additional extraction of SDOH, treatment regimen type (mono- vs. multi-therapy), and temporal features (e.g., start time of I/O).
* Foundation for building predictive models to estimate patient response likelihood and risk of adverse events in immunotherapy.

**Conclusion:** This project will demonstrate how LLMs can be applied to unstructured EHR data to extract high-value clinical information—especially immune-related adverse events—with minimal manual input. By enabling efficient and scalable data extraction, this work contributes to more personalized immunotherapy treatment strategies and advances the research goals outlined in Aim 1 of the larger study.

**Capstone Project Objectives and Deliverables:**

**Objectives:**

1. **Develop an NLP pipeline** using Large Language Models (LLMs), such as ClinicalBERT, to extract clinically meaningful entities from unstructured clinical notes in the IO Registry.
2. **Automate the identification and extraction of immune-related adverse events (irAEs)** to support downstream research and reduce reliance on manual curation.
3. **Extract additional contextual information** such as social determinants of health (SDOH), first-line therapy, treatment start times, and mono- vs. multi-therapy regimens.
4. **Enable time-series analysis** by linking extracted features to treatment timelines.
5. **Lay the groundwork for predictive modeling** of patient response to immunotherapy and the likelihood of adverse events.

**Deliverables:**

* A functioning **LLM-based NLP pipeline** for extracting irAEs and treatment-related features from clinical notes.
* A **labeled dataset** or evaluation framework comparing NLP output to a subset of ground truth annotations from clinical experts.
* **Documentation** of the methodology, model architecture, training data (if applicable), and evaluation metrics.
* A **summary report** detailing findings, limitations, and recommendations for future model improvement or deployment.
* A **presentation** to visualize extracted data and insights for clinical or research audiences.

**How many hours per week do you anticipate to work on this project: \_\_**20**\_\_\_**

*The HIDS Capstone project is a required course that has an experiential learning component. Each capstone project has been carefully reviewed and vetted by the MS Health Informatics and Data Science (HIDS) program to ensure that it meets the course requirements. The students will be collaboratively mentored by both a HIDS faculty member and Industry/Government co-mentor. The student will complete all the tasks on the Capstone project remotely or from an on-campus location. The student will not be paid for their work during the duration of the capstone project.*

**Approved by,**

**Adil Alaoui, MS, MBA**

**Capstone Course Director**