VCU College of Engineering

25-334: Large Language Models (LLM) for data extraction from clinical notes

Project Proposal

Prepared for

Rishabh Kapoor/Preetam Ghosh

VCU College of Engineering

By

Sawiya Aidarus, Connor Holden, Shashank Sinha, and August Moses

Under the supervision of

Preetam Ghosh

10/11/24

Executive Summary

This proposal outlines the Synthetic Note Generation and Large Language Model (LLM) Based Discrete Data Extraction project, which aims to address the challenge of extracting discrete data from free-text clinical notes. Current electronic medical records contain a wealth of unstructured data, but obtaining structured data (such as demographics, diagnosis, and treatment outcomes) from these notes is a manual, time-consuming process. The goal of this project is to build a robust tool for generating synthetic clinical notes, which can be used to train AI models for data extraction while avoiding issues related to protected health information (PHI).

This report covers the project's problem statement, engineering design requirements, scope of work, timeline, and team contract. Key deliverables include an expanded note generation system and a functional web tool for generating and customizing synthetic clinical notes. The milestones reflect a phased approach to development, and resources such as the Python-based system and available clinical note data are already in place.

With the successful completion of this project, we anticipate a significant improvement in the ability to train AI models for extracting data from clinical notes, while ensuring patient privacy and facilitating inter-facility collaboration.

Table of Contents

Section A. Problem Statement 4

Section B. Engineering Design Requirements 6

B.1 Project Goals (i.e. Client Needs) 6

B.2 Design Objectives 6

B.3 Design Specifications and Constraints 7

B.4 Codes and Standards 7

Standards 8

Codes 8

Section C. Scope of Work 10

C.1 Deliverables 10

C.2 Milestones 11

C.3 Resources 12

Appendix 1: Project Timeline 13

Appendix 2: Team Contract (i.e. Team Organization) 14

References 20

Section A. Problem Statement

The Synthetic Note Generation and LLM-Based Discrete Data Extraction Project addresses a critical gap in the field of radiation oncology, where clinical documentation is both time-consuming and prone to errors. Electronic Medical Record (EMR) systems store vast amounts of patient data, much of which exists as free-text clinical notes. These notes, though essential for documenting patient encounters, present significant challenges when extracting discrete data—such as treatment outcomes and key clinical metrics—that are crucial for research, registries, and patient care improvement.

Manual data extraction from free-text notes is not only labor-intensive but also error-prone, leading to inefficiencies in healthcare delivery and difficulties in scaling data extraction across large healthcare institutions. According to a systematic review by Lee et al. (2020), errors in clinical notes, such as mis recorded dose calculations or incorrect treatment site documentation, can result in serious consequences, including treatment delays, suboptimal care, and even legal implications. These errors often arise from time pressures, lack of standardized documentation, and poorly designed EMR systems.

The project's overarching goal is to revolutionize clinical note generation in radiation oncology by developing a method to generate synthetic clinical notes that resemble real patient data but contain no Protected Health Information (PHI). This approach allows for safe and efficient fine-tuning of Large Language Models (LLMs) to automatically extract discrete data, bypassing the risks of PHI exposure. The synthetic notes will also enable cross-institutional data sharing without compromising patient privacy.

Current efforts to automate clinical note generation and data extraction have been promising but remain limited in scope. For instance, the fine-tuned ClinicalBioBERT model has achieved a 95% weighted F1 score in extracting 27 discrete features from synthetic radiation oncology notes. However, this project aims to push beyond existing limitations by expanding the synthetic note generation tool, fine-tuning LLMs with a broader range of clinical notes, and developing a web-based application for customizable note generation.

This project addresses the unmet engineering need for more accurate and efficient clinical documentation and data extraction, with significant implications for enhancing patient care, reducing provider burden, and improving research outcomes in radiation oncology. By advancing the use of machine learning in healthcare, this project aims to establish new standards in clinical documentation practices.

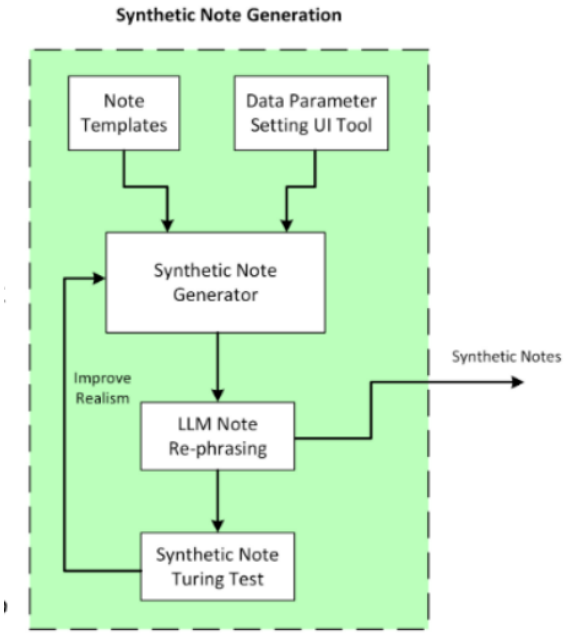

Figure 1: High level diagram of the note generation process

Figure 1. High Level diagram of the note generation process.

Section B. Engineering Design Requirements

This section describes the goals and objectives of the project, as well as all **realistic constraints** to which the design is bound. It is meant to provide a structure that helps to formulate the problem. Design requirements are often derived from client or stakeholder needs. They may consider benchmarking against or improving on currently available solutions, providing novel techniques or design solutions, integration with existing components, systems, or equipment, required codes and standards, general observations of the problem space, etc. Describe how the requirements provided below were researched and decided upon. Common design requirements often include considerations of the design efficacy, cost, safety, reliability, usability, and risk, among others.roposal, but is required for all subsequent reports. This section should be comprehensive and thorough, requiring a significant research effort.

B.1 Project Goals (i.e. Client Needs)

Our project follows a typical software design project including steps such as developing functional code, creating a web-based interface, and data preparation for testing and feedback. Each of these steps has one or more associated goals, however, the primary goal of our work is to enhance the functionality of the preexisting synthetic note-generator system.

- To create additional note templates to support multiple types of clinical notes and disease sites
- To parameterize LLMs to increase the accuracy of text-rephrasing and generation
- To implement a user-friendly web-tool that can be used by our clients
- To prepare data for a Clinical Turing Test to provide feedback for higher product accuracy
- To implement code that will generate a sequential episode of care

B.2 Design Objectives

Our project objectives are as follows:

- The design will consist of a larger variety of patient note templates that can be generated online with the help of the Groq API.
- The design will contain more note templates than previously
- The design will possibly utilize a small amount of project funds for our Groq API key.
- The design will have a prototype web-tool utilizing the updated synthetic note-generator by the end of the semester

B.3 Design Specifications and Constraints

Our current specifications are generally focused around updating existing code to generate a larger variety of data and creating a fully functional web-tool that allows users to change note generation parameters. While unsure of how many note templates or web parameters will be ideal, we have decided on a few numbers to aim for.

Project specifications are as follows:

- Design should have at least 5 *new* templates within the Synthetic Note Generator code
- Design should have a functional web-tool that allows users to adjust at least *3 parameters* for note generation, including choosing the note template itself

Our most unavoidable constraint is working with pre-existing code. While the existing code is currently helpful, it may produce roadblocks later and lead us to work around annoyances that cannot be changed. Other primary constraints include factors such as funding and computational costliness. It's essential for us to factor in the budget of our stakeholders and decide which elements they must pay for in the long run such as hosting a server or using an API subscription. The efficiency of our design can increase long-term costs in certain scenarios; therefore, we may need to format our design around an outside budget.

Project constraints include:

- Design must not let web users over-use Groq API since it is priced at tokens per second. Funds must not exceed $1,000 for our development process.
- Design for our web-tool must be user-friendly and allow for user specification.
- Design, if implemented locally, should work efficiently on a variety of machines such as Windows, Linux, and macOS.

B.4 Codes and Standards

The design and development of the Synthetic Note Generation Tool for clinical notes must adhere to several relevant codes and standards to ensure quality, safety, privacy, and compliance with best practices. These codes and standards will guide the technical and ethical aspects of the project, particularly in relation to web development, data privacy, and the handling of synthetic clinical data.

Standards

1. **IEEE Standard 1012-2016 – Systems and Software Verification and Validation**

This standard ensures that the software development process includes proper verification and validation steps, ensuring that the web tool functions as intended and meets the project's requirements. We will adopt this standard to guide our software testing and validation procedures, ensuring a reliable and bug-free user experience.

2. **W3C Standards for Web Development (HTML5, CSS, and JavaScript)**

The World Wide Web Consortium (W3C) sets global standards for web development to ensure accessibility, interoperability, and performance. Our web-based note generation tool will comply with the latest W3C standards for HTML5, CSS, and JavaScript to ensure cross-browser compatibility and adherence to best practices in web design. This will help ensure that the tool can be used across different platforms and devices without compatibility issues.

3. **ISO/IEC 27001:2013 – Information Security Management Systems (ISMS)**

Since the tool handles synthetic clinical data, it is essential to adhere to information security standards. ISO 27001 provides guidelines for maintaining security best practices, including risk management and access control. Though the data is synthetic and non-identifiable, this standard will ensure that our system is secure and protected from unauthorized access, safeguarding sensitive medical formats and structures.

4. **ACM Code of Ethics and Professional Conduct**

As members of the computing profession, our team will adhere to the ethical standards outlined by the ACM Code of Ethics. This involves ensuring that our design respects privacy, acknowledges data sensitivity, and promotes transparency in how synthetic notes are generated and managed. We will ensure that no synthetic notes resemble real patient data and maintain high ethical standards in the use of LLMs and data generation.

5. **NIST SP 800-53 – Security and Privacy Controls for Federal Information Systems and Organizations**

Since we are working with synthetic clinical notes and may need to interface with sensitive healthcare systems in the future, following the NIST guidelines ensures that our tool meets robust privacy and security controls. This is especially important as we design the system with privacy in mind, preventing PHI exposure even with synthetic data.

Codes

1. **HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule**

Although the tool generates synthetic notes, it must still respect the regulations outlined in HIPAA, ensuring that the system does not accidentally expose any real PHI (Protected Health

Information). Any data-handling process in the project must account for the legal and ethical constraints of HIPAA, particularly since our work could inform future developments involving real patient data.

2. **OSHA 1910.120 – Safety and Health Regulations for IT Environments**

While this code primarily deals with workplace safety, it is relevant to ensuring that the team follows safe IT practices while developing the tool. This includes ergonomic safety, electrical safety for computing environments, and proper handling of electronic components involved in the project.

Incorporating these codes and standards into the design ensures that the project aligns with industry best practices, legal requirements, and ethical standards. It also helps to guide the development process, ensuring that our software is secure, reliable, and designed with a focus on data integrity and privacy.

Section C. Scope of Work

The Synthetic Note Generation and LLM-Based Discrete Data Extraction project focuses on building a system for generating synthetic clinical notes and extracting discrete data using large language models (LLMs). This project will expand the existing note generation tool to support various types of clinical notes and disease sites, initially focusing on prostate and lung cancer. LLMs will be integrated to rephrase notes, adding variability while maintaining clinical accuracy. Additionally, a web-based interface will be developed to allow users to customize note generation parameters. The project will follow an Agile methodology, allowing for flexibility and continuous validation. Clear milestones and deliverables will be set to prevent scope creep, with regular communication with sponsors and advisors to ensure objectives are met on time and that the project stays within the budget. The key deliverables include the note generation system, an LLM-based extraction tool, and clinical validation through a Turing test.

C.1 Deliverables

In order to meet the project requirements given to the team by the sponsors, our team is choosing an Agile methodology to approach the project requirements. By setting specific deliverables at the end of each sprint, we will ensure continuous progress and flexibility in adapting to changes. This iterative approach allows for regular feedback from the project sponsor and faculty advisor, enabling us to refine the solution incrementally while minimizing risks. It also helps maintain clear communication, keep the project on track, and ensure that all objectives are completed on time and within budget.

- The synthetic note generator must be capable of handling various types of clinical notes. Currently, the system is limited to radiation oncology prostate consult notes, but it will need to expand to include other clinical note types used in cancer treatment, such as on-treatment visits, treatment summaries, and follow-ups.
- The synthetic note generation system will be expanded to support additional disease sites beyond the initial focus on prostate cancer. The system will incorporate lung cancer, requiring the creation of sub-templates tailored to lung cancer-specific clinical notes. Furthermore, the design will be flexible to accommodate more cancer types in the future, ensuring scalability and adaptability as the system evolves.
- To address the limitation of the static template system's lack of variability, we will incorporate pre-trained models like Groq to rephrase sections of the synthetic notes. This rephrasing will maintain the original meaning and preserve all discrete feature values, while adding variability to mimic the writing styles of specific medical providers or institutions. Users will also have control over the degree of rephrasing, enabling them to adjust the variability from minor tweaks to significant changes, ensuring consistency and flexibility across different healthcare settings.

- To ensure that the synthetic notes generated are truly indistinguishable from real clinical notes, we will conduct a Turing Test-style experiment. In this test, both real and synthetic notes will be presented to physicians, who will attempt to identify which notes are synthetic. If they correctly identify synthetic notes, we will gather feedback from them on any notes they have about what allowed them to figure out the notes were synthetic, using this insight to analyze and improve the note generation process.
- To make the synthetic note generator available more to the public, we will develop a web-based configuration tool that allows users to define the data ranges they wish to include and select specific disease sites to target. This will make the system more extendable and user-friendly, giving users greater control over the customization of generated notes.
- The system should also be able to generate a series of sequential notes following a cancer patient's care, from initial consultation through treatment and recovery. Starting with demographics, diagnosis, and care plans, it will create on-treatment notes documenting vitals, side effects, and responses to therapy. After treatment, follow-up notes will summarize recovery and long-term care. Randomized values and rephrasing will ensure the notes form a realistic, coherent narrative while maintaining clinical accuracy.

Almost all of the deliverables identified above can be done remotely, as all of them require research on the team's side, while being able to effectively collaborate with the sponsors to ensure all the specific requirements are achieved. For specifically the Turing Test-style experiment, this will require interactions with different physicians, either virtually or in-person, in order to gather notes about their view on the synthetic or clinical notes shown to them.

C.2 Milestones

Milestones	Expected Completion Date	Summar of milestone
Create the Web Based Configuration Tool	November 2, 2024	This task involves creating the web interface that users will interact with in order to create synthetic notes where they can specify the data ranges they want to include in their notes and what kind of disease sites to target.
Incorporate LLM to Rephrase Text	October 26, 2024	Coming up with the proper prompt for an LLM, such as Groq, such that the model can rephrase yet maintain the original meaning of a clinical note and the different sections.

Support an Additional Disease Site	October 26, 2024	Currently, the system only covers prostate cancer and will require to include at least lung cancer. This will require additional templates for lung cancer and should be designed in a way to handle more cancer types in the future.
Prepare data to create a Clinical Turing-test tool	October 26, 2024	In order to validate the synthetic clinical notes that are being generated, we want to be able to see if physicians, while being shown a real and synthetic note, are able to tell the difference between the two. We want to determine what method will be most effective in asking physicians these questions, and using the feedback in order to update the current clinical note templates.

C.3 Resources

In order to successfully complete the Synthetic Note Generation and LLM-Based Discrete Data Extraction project, several resources will be required. The Groq platform will be used for paraphrasing synthetic clinical notes, and will involve future costs for integrating it into the final web interface. To develop the web tool, frameworks such as Flask or Django will be necessary for the backend, along with front-end libraries like React to ensure a user-friendly interface. Cloud hosting services, such as AWS, Google Cloud, or Azure, may be needed in the future in order to deploy the interface for public access. A robust database system, either a relational or non-relational database, will be required to store generated notes and user inputs. Additionally, APIs for Groq integration and machine learning libraries for model fine-tuning will be critical. Version control systems like Git and GitHub will also be essential to manage the codebase and facilitate team collaboration. These resources will ensure the project meets its goals efficiently.

Appendix 1: Project Timeline

Gantt Chart of project timeline:

https://view.monday.com/embed/7620149158-8f2e43d20a3c65545058a163c0c870d5?r=use1

<iframe src="https://view.monday.com/embed/7620149158-8f2e43d20a3c65545058a163c0c870d5?r=use1" width=770 height=500 style="border: 0; box-shadow: 5px 5px 56px 0px rgba(0,0,0,0.25);"></iframe>

Appendix 2: Team Contract (i.e. Team Organization)

Step 1: Get to Know One Another. Gather Basic Information.

Task: This initial time together is important to form a strong team dynamic and get to know each other more as people outside of class time. Consider ways to develop positive working relationships with others, while remaining open and personal. Learn each other's strengths and discuss good/bad team experiences. This is also a good opportunity to start to better understand each other's communication and working styles.

Team Member Name	***Strengths each member bring to the group***	***Other Info***	***Contact Info***
Sawiya Aidarus	Good with organization, good with communication, and planning things strategically.	*I'm looking forward to learning new skills and creating a project with my peers.*	Aidarussa@vcu.edu 612-478-2257
Connor Holden	Good with organization, quick learner, and acquainted with professional environment.	*I'm excited to learn more about LLMs and apply my experiences.*	Holdencj@vcu.edu 571-287-3963
Shashank Sinha	Very flexible, can manage time very well, and previous experience.	*I'm looking forward to learning new skills and working on a real-world project.*	Sinhas6@vcu.edu 757-271-2877
August Moses	Previous experience with topic, being able to outline project deliverables and meet them.	*I'm looking forward to learning how to meet the needs of the stakeholder.*	Mosesa3@vcu.edu 540-645-8564

Other Stakeholders	***Notes***	***Contact Info***
Faculty Advisor: Preetam Ghosh		Pghosh@vcu.edu
Sponsor: Rishabh Kapoor		Rishabh.kapoor@vcuhealth.org

Step 2: Team Culture. Clarify the Group's Purpose and Culture Goals.

Task: Discuss how each team member wants to be treated to encourage them to make valuable contributions to the group and how each team member would like to feel recognized for their efforts. Discuss how the team will foster an environment where each team member feels they are accountable for their actions and the way they contribute to the project. These are your Culture Goals (left column). How do the students demonstrate these culture goals? These are your Actions (middle column). Finally, how do students deviate from the team's culture goals? What are ways that other team members can notice when that culture goal is no longer being honored in team dynamics? These are your Warning Signs (right column).

Resources: More information and an example Team Culture can be found in the Biodesign Student Guide "Intentional Teamwork" page (webpage | PDF)

Culture Goals	***Actions***	***Warning Signs***
Being on time to every meeting	- *Set up meetings reminders in discord* - *Send reminder e-mail the day before meeting*	- *Student misses first meeting without notice, warning is granted* - *Student misses' multiple meetings afterwards – issue is brought up with faculty advisor*
Informing the group of any delays in completing assignments	- *Weekly progress check during student meetings* - *Set reasonable deadlines and note when an extension is needed*	- *Student shows up for weekly meeting with no considerable work done, then a discussion would need to be had* - *Student shows up for weekly meeting with no considerable work done on multiple occasions – issue is brought up with faculty advisor*
Have a good balance of work between members	- *Check weekly at meetings that every has a fair workload, make sure everyone agrees*	- *Issues should be brought up during meetings and we can disperse work accordingly*

		- *Whoever feels like they aren't being heard after communicating, they can bring up issues during meeting with faculty advisor*

Step 3: Time Commitments, Meeting Structure, and Communication

Task: Discuss the anticipated time commitments for the group project. Consider the following questions (don't answer these questions in the box below):

- What are reasonable time commitments for everyone to invest in this project?
- What other activities and commitments do group members have in their lives?
- How will we communicate with each other?
- When will we meet as a team? Where will we meet? How Often?
- Who will run the meetings? Will there be an assigned team leader or scribe? Does that position rotate or will same person take on that role for the duration of the project?

Required: How often you will meet with your faculty advisor, where you will meet, and how the meetings will be conducted. Who arranges these meetings?
See examples below.

Meeting Participants	***Frequency*** ***Dates and Times / Locations***	***Meeting Goals*** ***Responsible Party***
Students Only	*Every other week on Thursday (6pm-7pm), On Discord Voice Channel*	*Update group on day-to-day challenges and accomplishments (Connor will record these for the weekly progress reports and meetings with advisor)*
Students Only	*Every other week on Thursday (6pm-7pm), in ENGR West 0101*	*Actively work on project (August will document these meetings by taking photos of whiteboards, physical prototypes, etc, then post on Discord and update Capstone Report)*
Students + Faculty advisor	*Every Thursday at 5 pm in Advisor's office (conference room if they can reserve)*	*Update faculty advisor and get answers to our questions (Connor will scribe; Sawiya will create meeting agenda and lead meeting)*

Project Sponsor	*First Friday at noon of every month* *If sponsor is available, we'll figure out Zoom or in person details* *If not, then we'll update the sponsor via email.*	*Update project sponsor and make sure we are on the right track (Connor will scribe; Sawiya will create meeting agenda and lead meeting; Shashank will present prototype so far)*

Step 4: Determine Individual Roles and Responsibilities

Task: As part of the Capstone Team experience, each member will take on a leadership role, ***in addition to*** contributing to the overall weekly action items for the project. Some common leadership roles for Capstone projects are listed below. Other roles may be assigned with approval of your faculty advisor as deemed fit for the project. For the entirety of the project, you should communicate progress to your advisor specifically with regard to your role.

- **Before meeting with your team**, take some time to ask yourself: what is my "natural" role in this group (strengths)? How can I use this experience to help me grow and develop more?
- **As a group,** discuss the various tasks needed for the project and role preferences. Then assign roles in the table on the next page. Try to create a team dynamic that is fair and equitable, while promoting the strengths of each member.

Communication Leaders

Suggested: Assign a team member to be the primary contact for the client/sponsor. This person will schedule meetings, send updates, and ensure deliverables are met.

Suggested: Assign a team member to be the primary contact for faculty advisor. This person will schedule meetings, send updates, and ensure deliverables are met.

Common Leadership Roles for Capstone

1. **Project Manager:** Manages all tasks; develops overall schedule for project; writes agendas and runs meetings; reviews and monitors individual action items; creates an environment where team members are respected, take risks and feel safe expressing their ideas.
   Required: On Edusourced, under the Team tab, make sure that this student is assigned the Project Manager role. This is required so that Capstone program staff can easily identify a single contact person, especially for items like Purchasing and Receiving project supplies.
2. **Logistics Manager:** coordinates all internal and external interactions; lead in establishing contact within and outside of organization, following up on communication of commitments, obtaining information for the team; documents meeting minutes; manages facility and resource usage.
3. **Financial Manager:** researches/benchmarks technical purchases and acquisitions; conducts pricing analysis and budget justifications on proposed purchases; carries out team purchase requests; monitors team budget.

4. **Systems Engineer:** analyzes Client initial design specification and leads establishment of product specifications; monitors, coordinates and manages integration of sub-systems in the prototype; develops and recommends system architecture and manages product interfaces.
5. **Test Engineer:** oversees experimental design, test plan, procedures and data analysis; acquires data acquisition equipment and any necessary software; establishes test protocols and schedules; oversees statistical analysis of results; leads presentation of experimental finding and resulting recommendations.
6. **Manufacturing Engineer:** coordinates all fabrication required to meet final prototype requirements; oversees that all engineering drawings meet the requirements of machine shop or vendor; reviews designs to ensure design for manufacturing; determines realistic timing for fabrication and quality; develops schedule for all manufacturing.

Team Member	***Role(s)***	***Responsibilities***
Sawiya Aidarus	*Project Manager*	- *Send out weekly emails and other correspondence* - *Create schedule and meetings* - *Make sure everyone understands what is going on*
Connor Holden	Logistics Manager	- *Keep a detailed record of meeting notes and share with group* - *Coordinate meeting times* - *Follow up with communication of commitments* - *Obtaining information for the team* - *Manages facility and resource usage*
Shashank Sinha	Financial Manager	- *Keeps track of team budget* - *Sends in request for any additional money* - *Conducts pricing analysis*
August Moses	Test Engineer	- Oversees testing and implementation - Acquires equipment and any necessary software - Oversees statistical analysis of results

Step 5: Agree to the above team contract

Team Member: Sawiya Aidarus *Signature:* _____

Team Member: Connor Holden *Signature:* _____

Team Member: Shashank Sinha *Signature:* _____

Team Member: August Moses *Signature:* _____

References

Provide a numbered list of all references in order of appearance using APA citation format. The reference page should begin on a new page as shown here.

[1] VCU Writing Center. (2021, September 8). *APA Citation: A guide to formatting in APA style.* Retrieved September 2, 2024. https://writing.vcu.edu/student-resources/apa-citations/

[2] Teach Engineering. *Engineering Design Process*. TeachEngineering.org. Retreived September 2, 2024. https://www.teachengineering.org/populartopics/designprocess

[3] Lee, J., et al. (2020). Errors in clinical notes in radiation oncology: A systematic review. *International Journal of Radiation Oncology, Biology, Physics, 106*(3), 531-538.