DATA RECIPIENT AGREEMENT
FOR DE-IDENTIFIED DATASET

This Data Recipient Agreement ("Agreement") with respect to a De-Identified Data Set is entered into by and between **University of Massachusetts Medical School** (hereinafter referred to as "Data Provider"), a public institution of higher education within the Commonwealth of Massachusetts, 55 Lake Avenue North Worcester MA 01655 and [Recipient] (hereinafter referred to as "Data Recipient"). The Agreement is entered into as of the date of last required signature on the agreement.("Effective Date").

RECITALS

Data Provider and Data Recipient desire to set forth the terms and conditions under which Data Provider will disclose to Data Recipient certain De-Identified Health Information (De-identified Data Set or Data Set) described in this Agreement for the research described in this Agreement.

In consideration of the mutual promises below, and Data Provider's disclosure of the Data Set to Data Recipient under this Agreement, the parties agree as follows:

ARTICLE I
DEFINITIONS

1.1 **De-Identified Data,** as defined in the Privacy Rule at 45 CFR Section 164.514(a) and (b), as may be amended.

1.2 **Security Rule** means the Standards for Security for the Protection of Electronic Protected Health Information, codified at 45 CFR parts 160 and 164, Subpart C, effective April 20, 2005, as may be amended.

1.3 The following terms shall also have the meanings given to them in the Privacy Rule: Covered Entity, Individual, Protected Health Information, and Required by Law.

ARTICLE II
DATA PROVIDER'S OBLIGATIONS

2.1. Data Provider will disclose to Data Recipient the De-Identified Data Set for the purposes of the research study entitled "Developing Natural Language Processing and Information Retrieval Technology for Electronic Health Records (EHR) to Improve Patient Education, Patient-Provider Communication, and Clinical Knowledge Discovery". The study is more specifically described in Exhibit A ("the Study"). Exhibit A, which also describes the De-Identified Data Set, is attached to, and by this reference incorporated in, this Agreement.

2.2. Data Provider shall not request Data Recipient to use or disclose the De-Identified Data Set in any manner that would violate the Privacy Rule or applicable laws or regulations.

ARTICLE III
DATA RECIPIENT'S OBLIGATIONS

3.1. Unless specifically stated otherwise in this Agreement, Data Recipient's obligations with respect to the De-Identified Data Set apply to the whole and to any part of the Data Set.

3.2. Data Recipient shall not use or disclose the Data Set for any purpose other than the Study as authorized by this agreement or as Required by Law. In addition, Data Recipient shall not use or disclose the Data Set in any manner that would violate the Privacy Rule or applicable laws or regulations.

3.3. Exhibit A specifies who is permitted to use or receive the Data Set for the purposes of the Study.

3.4. Data Recipient may not subcontract its performance obligations, or assign its rights, under this Agreement without the express written consent of Data Provider. Data Recipient shall ensure that any subcontractor agrees in writing to the same terms and conditions regarding a Data Set that apply to Data Recipient under this Agreement.

3.5. Data Recipient must have appropriate safeguards to prevent the use or disclosure of the Data Set in any manner not permitted by this Agreement.

3.6. Data Recipient must not identify or contact (or attempt to do so) either directly or through another person, any Individual in the Data Set.

3.7. Data Recipient agrees to mitigate, to the extent feasible and allowed by law, any harmful effect that is known or becomes known to Data Recipient that arises from a use or disclosure of the Data Set by Data Recipient or its agents in violation of this Agreement or the Privacy Rule.

3.8. Data Recipient must notify Data Provider within twenty-four (24) hours by phone, and in writing within five (5) business days, after Data Recipient becomes aware of any use or disclosure not authorized by the Agreement and any actual or suspected breach of Data Recipient's security.

3.9. Data Recipient acknowledges that Data Recipient has no ownership rights in the Data Set.

3.10. Within ten (10) business days of a written request by Data Provider, Data Recipient shall allow Data Provider to conduct a reasonable inspection, at a mutually agreeable time, of Data Recipient's facilities, systems, books, records, agreements, and

policies and procedures relating to the use or disclosure of the Data Set for the purpose of determining Data Recipient's compliance with this Agreement. Any failure of Data Provider to inspect or to detect or notify Data Recipient of an unsatisfactory practice does not constitute acceptance of the practice by Data Provider or a waiver of any remedy or right Data Provider has under the Agreement or applicable law.

3.11 Standards for Security of Data Set. To the extent that Data Recipient creates, receives, maintains, or transmits information, Data Recipient shall also have administrative, physical and technical safeguards that reasonably and appropriately protect the confidentiality, integrity, and availability of any electronic Information that may be transmitted in conformity with the requirements of the Security Rule and other appropriate and applicable requirements.

3.12 Reporting of Security Incidents. If the Data Recipient creates, receives, maintains, or transmits electronic PHI, Data Recipient shall appropriately report any incident, as defined by the Security Rule.

3.13 Mitigation of Security Incidents. Data Recipient shall mitigate promptly, to the extent practicable, any harmful effect that is known to Data Recipient caused by a security incident regarding electronic Protected Information by Data Recipient in violation of this Agreement, the Security Rule, or other applicable federal or state law.

3.14. Data Recipient shall comply with state security and privacy laws to the extent that they are more protective of the Individual's privacy than the HIPAA Privacy Rule.

3.15 To the extent permitted by law, Data recipient will indemnify Data Provider from and against any and all claims, losses, liabilities, costs and other expenses (including attorneys fees) that result from or arise directly or indirectly out of or in connection with any negligent act or omission or willful misconduct of Data Recipient, its officers, employees, agents or subcontractors relative to the Data Set, including without limitation, any violation of Data Recipient's responsibilities under this Agreement with respect to the Data Set.

ARTICLE IV
AMENDMENT AND TERMINATION

4.1 When Data Provider reasonably concludes that an amendment to the Agreement is necessary to comply with applicable law, Data Provider shall notify Data Recipient in writing of the proposed modification(s) ("Legally-Required Modifications"). Data Provider shall request Data Recipients written approval in the form of an amendment to this agreement at the time of notification. Data Recipient shall have thirty (30) days to sign the amendment and return it to Data Provider. Data Recipient's rejection of a Legally Required Modification is grounds for termination of the Agreement by Data Provider on thirty (30) days written notice.

4.2. A breach by Data Recipient of any provision of the Agreement, as determined by Data Provider, shall constitute a material breach and grounds for immediate termination of the Agreement by Data Provider. At its sole discretion, Data Provider may give Data Recipient 30 days to cure the breach.

4.3. On termination of the Agreement for any reason, Data Recipient shall return or destroy the Data Set. If return or destruction is not feasible, Data Recipient shall explain to Data Provider why, in writing, to the address given in this Agreement.

4.3.1. If Required by Law, Data Recipient may retain documentation for the time specified as necessary to comply with the law.

4.3.2. Data Recipient's obligations under this Agreement shall continue until Data Recipient destroys the Data Set or returns the information to Data Provider; provided however, that on termination of the Agreement, Data Recipient shall not further use or disclose the Data Set except as Required by Law.

4.4 If Data Recipient elects to destroy the Data Set, Data Recipient shall certify in writing to Data Provider that the Data Set has been destroyed.

ARTICLE V
MISCELLANEOUS

5.1. Exhibit A may be modified by the parties at any time pursuant to a writing executed by both parties. No use or disclosure different from that permitted by the currently in force Exhibit A may be made until the new Exhibit A has been signed by both parties.

5.2. Any ambiguity in this Agreement relating to the use and disclosure of the Data Set by Data Recipient shall be resolved in favor of a meaning that further protects the privacy and security of the information.

5.3 All notices required or permitted under the Agreement to be in writing may be delivered personally, by electronic facsimile (with a confirmation by registered or certified mail placed in the mail no later than the following day), or by registered or certified mail, postage prepaid, addressed to a party as indicated below:

If to Data Provider:	If to Data Recipient:
Margaret (Meg) Johnson, JD, CIP Director, Office of Clinical Research University of Massachusetts Worcester Phone: (508) 856-5152 Facsimile No.: (508) 856-1980 Email: Meg.Johnson@umassmed.edu	[Name] [Title] [Institution] [Address] Phone: Facsimile No.: Email:

Notice shall be deemed to have been given on receipt of communications personally delivered or transmitted by electronic facsimile (delivery confirmed) and, for communications made by United States mail, on the third (3rd) day after mailing. The above addresses may be changed by giving written notice as described in this section.

5.4. Data Recipient's obligations under Article IV of this Agreement shall survive the termination of the Agreement.

5.5 If any provision of this Agreement is determined by a court of competent jurisdiction to be invalid, void, or unenforceable, the remaining provisions shall continue in full force and effect.

5.6. This agreement shall be construed and enforced under the laws of the Commonwealth of Massachusetts

IN WITNESS WHEREOF, the parties agree as follows:

Data Provider	**Data Recipient**
By: X _____ PI of Data Provider Group	By: X _____ Leader of Data Recipient Group
Name: Hong Yu, PhD	Name:
Title: Professor, Department of Quantitative Health Sciences University of Massachusetts Medical School	Title:
Date: 10/25/2017	Date:
By: GJCampbell	X _____ Privacy Officer, Legal Counsel, ect
Name: Gerry Campbell, JD MS	Name:
Title: Senior Privacy Officer	Title:
Date: 7/20/17	Date:

EXHIBIT A

To the Data Agreement for Research
Study Title: Developing Natural Language Processing and Information Retrieval Technology for Electronic Health Records (EHR) to Improve Patient Education, Patient-Provider Communication, and Clinical Knowledge Discovery

1. **Description of the Study:**
   Objectives*

   Develop Natural Language Processing (NLP) and information retrieval (IR) approaches to automatically process Electronic Health Records (EHRs), including identifying medical concepts (or jargon) and concept relations, translating medical jargon into lay language or a language other than English (starting with Spanish), linking EHR records to definitions and other education materials, and generating EHR-specific clinical questions for both patients and physicians.

   NLP and IR
   In order to develop, innovate and iteratively improve NLP and IR approaches, we will automatically process all EHRs at UMass Memorial Health Care (UMMHC), for the purpose of: medical concept identification, concept prioritization, relation identification (that is the semantic relations, including causal and temporal relations, between two or more concepts), feature generation and NLP and IR algorithm improvement. We will develop advanced technologies including deep learning, word embedding, semantic/discourse analysis of EHR content, text summarization and machine learning.

   All NLP and IR processes will be run as batch processes within the UMMHC secure firewall on EHRs inclusive of their identifiers. The outputs include ranked EHR notes, medical concepts, features and computational algorithms. We will remove any identifiable information if we make the NLP and IR output explicit (including conference presentations, group discussions, and publications).

   Corpora
   We will use manual annotation of EHR notes as training data for automated methods. We will manually annotate EHR notes to identify important medical concepts and clinical questions, to link EHR notes to external education materials, and to translate EHR notes to lay language and other languages (Spanish). All EHR notes will be de-identified (first automatically using a de-identification tool, followed by manual de-identification to ensure 100% de-identification). Prior to dissemination of any training corpus, we will modify the IRB docket and gain written permission from the IRB.

   Background*
   EHRs are a rich clinical resource from which we can develop advanced NLP and IR to benefit healthcare outcomes. For example, effective education and providing

patients with clinically-relevant information have been shown to influence patient behavior and produce the changes in knowledge, attitudes, and skills necessary to maintain or improve self-management. Traditional approaches have relied upon office-based face-to-face education for patient–physician communication. However, a "practice gap" exists such that patients, especially those with complex diseases such as diabetes, require more information, while physicians are increasingly limited by time constraints. Although other education aids, such as written handouts, visual aids, audiovisuals, and Internet materials, are useful adjuncts to patient education, they may not be tailored to information specific to a patient's clinical condition(s).

The Affordable Care Act places significant emphasis on patient centered care and outcomes; and there has been a strong push by the Agency for Healthcare Research & Quality (AHRQ) to encourage patient engagement through asking questions. The AHRQ has published a list of questions which patients should ask their doctors. The list includes questions such as "what is the best hospital for my needs?" While these questions are potentially useful, the generic nature of the question reduces the utility of the answer for the patient; and while asking questions is a significant part of the shared decision making process, a number of barriers exists. For example, vulnerable patients are less likely to ask questions even though they have the most benefit to gain from engaging with providers, physicians may find it hard to provide information at the level in which the patient is comfortable.

In addition, enhancing patients' access to their own clinical notes in their EHR has been seen as a central component of patient-centered care and is becoming increasingly common. However, a recent study involving over ten thousand patients has shown that patients—especially in vulnerable groups (e.g., lower literacy, lower income)—can be confused by EHR notes.

We know EHRs present opportunities for novel and personalized communication with the potential to increase each patient's involvement in their own care and to improve each patient's communication with their physicians and other caregivers. Innovative tools are needed to help patients understand and effectively use their clinical notes, and NLP approaches have been shown to improve text comprehension. We are developing multi-module NLP and IR systems. For example, one system called NoteAid will link clinical notes with medical jargon to definitions and related educational material from trusted resources. We will translate medical jargon into lay concepts and eventually translate clinical notes into lay term notes. We aim to personalize education materials based on automated analysis of patients' longitudinal EHRs.

We expect that our NLP and IR systems will improve patient comprehension of their EHR notes. We also expect it to help bridge the communication gap and enhance patient/physician dialog. These in turn will increase patient autonomy and disease self-management.

EHR NLP systems can also be developed for pharmacovigilance and drug surveillance. Adverse drug events (ADEs) are common, leading to adverse healthcare outcome and death in severe cases. Studies have shown that ADEs are described in EHRs. Manual abstract is prohibitively expensive and NLP systems that automatically identify ADEs will improve patient drug safety.

We will analyze data from patients with a spectrum of diseases but will begin with a focus on:

- Cardiac and cancer patients because of the scope of the problem
- Diabetic patients; diabetes is an epidemic that disproportionately affects minorities and the underserved.
- Diabetic and cardiac Hispanic (primarily Spanish speaking) patients

Inclusion and Exclusion Criteria*

We will use data collected in near real time, data that we collected on previous IRBs, and data from collaborators. All data sets will include records from children and pregnant women. We will exclude any data that we know to be from prisoners.

NEW DATA COLLECTION (new IRB)

Development and implementation EHR NLP and IR technology uses prospective and retrospective EHR records. It is prospective in that we will periodically collect new data which may include data subsequent to the IRB approval date. However, we are not tracking patients going forward in time or watching for an outcome. Furthermore, all data exists at the time of collection. Data collection will include records from children and pregnant women. We will exclude any data that we know to be from prisoners.

Use of UMASS identified data:

NLP and IR: All UMMHC prospective and retrospective patients EHR data to generate a sufficiently large dataset for development of NLP and IR. For example, we will collect HL7 messages to collect unstructured clinical narratives not currently found in the UMMS Clinical Data Repository. HL7 (Health Level Seven) is a standard for the exchange, integration, sharing, and retrieval of electronic health information. These standards define how information is packaged and communicated from one party to another, setting the language, structure and data types required for seamless integration between systems. Information sent using the HL7 standard is sent as a collection of one or more messages, each of which transmits one record or item of health-related information.

- Corpora: All UMMHC prospective and retrospective patients EHR data needed to complete data sets. For example, corpora with longitudinal clinical narratives in specific disease areas for disease specific clinical features or corpora for other modules such as translation or question generation).

Use of External data

- Use of external identified data: We will use identified data from other organizations. The co-investigator or collaborator will have their own institutional IRB approval, be responsible for data identification and security and we will comply with any data use agreements. For example in our work with Partners HealthCare, they use the same criteria for patient selection. We will remotely access their prospective and retrospective EHR data behind their firewall using encryption software and no data will leave their servers. All computational work will be performed on their servers and identified results will stay there. This work will be the same as for UMASS (NLP and IR/Corpora).

Use of external de-identified data:

- We will use de-identified data from other organizations. The co-investigator or collaborator will: have their own institutional IRB approval, be responsible for data de-identification and data transfer and will provide in writing that they will never break the code for a crosswalk or enable re-identification. We will comply with any data use agreements. For example in our work with Northwestern, they will select patients using the same criteria as we do and provide de-identified EHR data in a system agreed to by UMASS and Northwestern's IT groups. A description of the IT environments is in Sections 12 and 13. This work will be the same as for UMASS (NLP and IR/Corpora).
We will also do a manual review of their data that is de-identified using automated methods and work with them to improve PHI detection algorithms. De-identified data being manually annotated and reviewed for PHI (using SafeHarbor) will be housed in the same location as our UMMS data with PHI – the HIPAA compliant, regulated environment - and will stay there until fully reviewed. Fully de-identified notes and extracted data without PHI may be moved to the servers with controlled access.

USE OF PREVIOUSLY COLLECTED UMASS DATA (previous IRB approved data)

In order to minimize data access from UMMHC, we will also reuse UMASS data previously accessed with IRB approval provided it is consistent with any data use agreements. This data may include records from children and pregnant women. We will exclude any data that we know to be from prisoners.

NLP and IR:

- All UMMHC prospective and retrospective patient data in the dedicated patient data systems for "near real time" data access available to the program through IRB docket H00001387. This will be used for improvement of NLP tools for detection of clinical features in clinical narratives requires large scale data sets. This will also be used to detect ADEs from EHRs.

Corpora

- All UMMHC prospective and retrospective patient EHRs used for annotation of clinical narratives (for 2500 clinical notes each of diabetes, cardiovascular and cancer patients) available through IRB Docket H00001387, 200 EHRs from diabetes and cardiovascular patients (data for 75 patients has already been collected with consent under former related IRB Docket H00003068) and 100 EHRs from Spanish speaking diabetic patients (former IRB Docket H00003366). This data will be used as training data for supervised and semi-supervised machine learning algorithm development.

Study Endpoints*

1. Automated methods to allow patients to understand and use their own EHR notes by simplifying, translating, and offering suggested questions.
2. Detection of known and unknown ADEs from EHRs.

Procedures Involved*

The EHR data inclusive of identifiers will be aggregated for batch processing. It is important to maintain identifiers in the data set because:

- Endpoints of the research include pharmacovigilance and adverse event detection on real EHR data in real time. Use of original EHR data most closely approximates these conditions
- De-identification process introduces "noise" into the data and impacts the NLP results
- Current work includes improving algorithms for de-identification

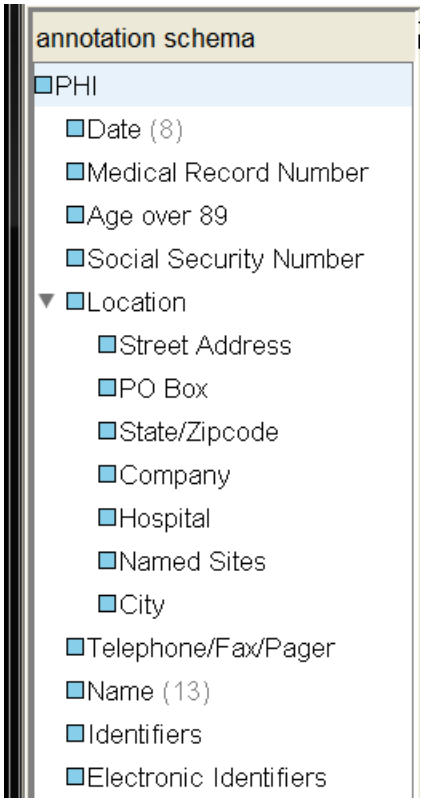
We will train NLP and IR models, including distributional representation, supervised, semi-supervised and unsupervised machine learning and deep learning, for iterative improvement of NLP and IR tools and for detection of clinical features in clinical narratives.

We will use de-identified EHR notes or reuse de-identified annotated notes. Annotation will include manually identifying and mark medical concepts. It will also involve translation of these concepts or jargon into lay language or other languages. Annotators will identify concepts important to link to educational materials or for inclusion in question/answering modules. The data will be used to create a corpus and inform supervised machine learning methods. The output will be improved NLP algorithms for identification of medical jargon and simplification of EHR notes with lay language and links to educational materials and question/answering content.

EHR data de-identification is done by Safe Harbor methods, specific processes are described below:

- EHR data is processed with modified "De-Identification V1.1" (http://www.physionet.org/physiotools/deid/) to identify each of the 18 types of Safe Harbor identifiers.
- It pre-annotates the clinical notes in light blue.

- Each clinical note is then manually reviewed to ensure all PHI is marked fully and correctly.
- Missing or incorrect annotation is annotated or fixed.
- Shown in the Figure below are the entities in our annotation schema and all 18 elements map to the 15 that are displayed. For example "Identifiers" is used to capture all non-electronic or medical record identifiers and may include serial numbers, vehicle licenses, etc. Medical record will cover medical record numbers, health plan numbers, professional license numbers, etc.
- If PHI is not pre-annotated, it is annotated in our tool, a modified Protégé with Knowtator plugin
- In addition to annotator marking PHI, an editor will review the annotators work.
- Annotation tags the PHI for computationally removing PHI or replacing it with fictitious information.
- Using this three step process is considered both thorough and defensible by QMS.



Descriptions of these procedures are found in the documents:

- "Anticoagulant ADE Research Strategy R01HL125089" in section D2 starting on page 7, especially D2.1 through D2.3 for model training and section D2.4 for evaluation
- "Cancer ADE Research Strategy U01CA108975" in section D2 starting on page 7, especially D2.1 through D2.3 for model training and section D2.4 for evaluation

- Also see Section 5 Data Requirements found in this document and refer to the above paragraph

2. **Data Set 1:**

   Data Set 1 comprises of a total of 1154 de-identified English EHR notes from cancer patients. Each note was annotated by two annotators who label medical entities into two groups: medical events and attributes. Medical event categories are Adverse Drug Event (ADE), Drug Name, Indication and Other Sign Symptom and Disease. Route, Frequency, Duration and Dosage are attributes of Drug Name. Severity is an attribute to ADE and SSD.

3. **Permitted Uses.** Data Recipient may only use Data Set 1 for research. The data can't be shared with any other party who is not signed in this agreement without a new application from the other party.