AWARD NUMBER: DM190529

TITLE: Developing Natural Language Processing Algorithms to Medically Code the Clinical Notes in the Theater Data Medical Store

PRINCIPAL INVESTIGATOR: LCDR Andrew Olson

CONTRACTING ORGANIZATION: Naval Health Research Center
San Diego, CA

REPORT DATE: September 2021

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
September 2021

2. REPORT TYPE
Annual Report

3. DATES COVERED
05Aug2020-04Aug2021

4. TITLE AND SUBTITLE
Developing Natural Language Processing Algorithms to Medically Code the Clinical Notes in the Theater Data Medical Store

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DM190529

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
LCDR Andrew Olson; Zouris, James; D'Souza, Edwin;
Elkins, Trevor; MacGregor,

E-Mail: andrew.s.olson22.mil@mail.mil

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Outside the Department of Defense, natural language processing (NLP) strategies have been used with electronic health records (EHR) to increase information extraction from free text notes and structured fields, allowing access to much larger cohorts than previously possible. Current operational medical data is held in the Theater Medical Data Store (TMDS). Most of the medical information in TMDS is contained in unstructured text fields. The objective will be to automate the data-coding process into the injury diagnostic code groups, which are derived from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes. There are over 8 million records in the TMDS and there may be as much as 50% of the ICD-9-CM codes that are not completely or accurately coded. The accuracy of the data in the TMDS has never been quantified, largely because most has been captured without any medical billing concerns.

The study has developed a set of programming rules using NLP and machine learning (ML) (i.e., algorithms generated by automated learning from manually coded data), with eventual output that will represent human interpretation as much as possible. The coding algorithm models have been developed using pre-existing coded medical records from the Expeditionary Medical Encounter Dataset (EMED) housed at the Naval Health Research Center (NHRC). Experienced nurse staff are responsible for coding and validating all the EMED medical encounter records. The model will be trained on a subset of the EMED data and then tested on TMDS data that has been matched to the remaining EMED data.

15. SUBJECT TERMS
natural language processing, machine learning, clinical notes

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified	Unclassified	Unclassified	Unclassified	9	USAMRMC
					19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. **INTRODUCTION:**

    - This research will develop a set of programming rules or machine learning algorithms (i.e., algorithms generated by automated learning from manually coded data) to determine the principal ICD-9-CM code, which is the diagnosis code that is primarily responsible for the admission of the patient to the hospital. The objective will be to develop machine learning algorithms using Natural Language Processing (NLP) to develop a dataset called the Automated Clinical Encounter Repository (ACER). The ACER will primarily be used as supplemental repository. The baseline data will be obtained using pre-existing coded medical records. At present, the majority of the ICD-9-CM coded are trauma-related, specifically wounded-in-action, and as such, this proposal will only use the "Injury and Poisoning" codes (i.e., 800–999). The ACER algorithms will then be tested on the matched electronic health records in the Theater Medical Data Store to determine accuracy. Three different models will be developed and analyzed using the ICD-9-CM coding hierarchy: 1) the trauma and anatomical location subcategory diagnostic code groups; 2) the first three digits of the ICD-9-CM codes; and 3) the billable or actual ICD-9-CM diagnostic codes.

2. **KEYWORDS:** natural language processing, machine learning, coding, clinical notes

3. **ACCOMPLISHMENTS:**

    - **What were the major goals of the project?**

        - Specific Aim 1: Predict the ICD-9-CM code by subcategory

            - Specific Aim 1 will predict the ICD-9-CM subcategory for each principal diagnosis of the TMDS. The model will be trained on a subset of the EMED data and then tested on TMDS data that has been matched to the remaining EMED data. This aim will be the most achievable, as there are only 24 subcategories for ICD-9-CM codes 800-999. A complete listing of the subcategories are shown in Table 1.

        - Specific Aim 2: Predict the 3-digit ICD-9-CM codes within each subcategory

            - Specific Aim 2 will predict the principal 3-digit ICD-9-CM codes from 800-999 for each record of the TMDS. The model will be trained on a subset of the EMED data and then tested on TMDS data that has been matched to the remaining EMED data. The data will consist of around 150 3-digit ICD-9-CM codes. It is envisioned that certain ICD-9-CM codes will have significantly more data allowing for greater reliability of the results.

        - Specific Aim 3: Predict the billable ICD-9-CM codes within each subcategory

            - Specific Aim 3 will predict the principal billable ICD-9-CM code for each record of the TMDS. The model will be trained on a subset of the EMED data and then tested on TMDS data that has been matched to the remaining EMED data. This aim will consist of trying to decipher and code over 1,500 billable ICD-9-CM codes and also identify the codes for which the algorithms work best. The goal will be to ascertain which billable ICD-9-CM codes can be predicted with at least a 70–80% accuracy. We will most likely restrict our coding to the best performing billable codes.

        - Specific Aim 4: Predict the secondary and tertiary ICD-9-CM codes in the TMDS

            - Specific Aim 4 will attempt to the code additional ICD-9-CM codes from the TMDS in addition to the principal code. Patient records with secondary and tertiary ICD-9-CM will be selected and trained from the EMED data and then tested on the same record in the TMDS. This aim will be the most difficult as patients can have multiple codes per record but will restrict to the first three diagnoses of the patient (primary, secondary and tertiary).

        - Specific Aim 5: Development of the ACER using all the TMDS data

        - Specific Aim 5 will use the algorithms developed from Specific Aims 1–4 to develop the ACER. The goal is to identify and predict the coding for trauma patients present in the TMDS that have not been coded in the EMED, which will enable trauma patients to be expedited into the registry with considerably less clinical review. We will most likely restrict our coding to the best performing billable codes.

    - Specific Aim 6: Determine the accuracy of the principal ICD-9-CM codes in TMDS

        - The accuracy of the ICD-9-CM data in the TMDS has never been quantified, primarily because most of the data has been captured in Iraq and Afghanistan without any medical billing concerns. Sensitivity, specificity, and accuracy metrics will be compared for Specific Aims 1–5.

- **What was accomplished under these goals?**

    - Aim 1: 83% overall accuracy

    - Aim 2: 73% overall accuracy

    - Aim 3: 53% overall accuracy

        - Overall accuracy ~60% removing codes 850.9, 850.5, and 850.12

        - Over 30 billable ICD-9 codes had an accuracy of more than 50%

            - Represents 75% of the data

        - Exploring topic modeling and word vectors provided additional information but more work is needed

    - Removing common phrases prior to text preprocessing is very important in model performance

    - Document term matrix provides an excellent approach in estimating the ICD-9 subcategories

    - Inverse document frequency (tf-idf) Random forests and extreme gradient boosting were the best models

    - Deep Learning methods are not performing as well as 'shallow' learning models (Random Forests, XGBoost)

    - Only the Subjective Note was used instead of the SOAP note however as stated above the expanded SOAP notes didn't improve RF and XGBoost

    - There is not enough data to apply multilabel classification using billable ICD-9 codes

        - Need to use 3-digit ICD-9 codes

    - Aim4 results for Multi-label were not good primarily due to the lack of data however there are some positive takeaways

        - PPV was 78% (Given the predicted value is true how accurate is the model performing

        - Implementing and understanding the Multi-label Classification will be useful for future studies

    - Aim5 results are preliminary but potentially over hundred-thousand of records may be partially automatically coded

- Aim6 results are preliminary but potentially thousands of records may be partially automatically coded
  - There is a strong likelihood that the TMDS have a significant portion of misclassified coded diagnoses
  - Among the useable records in TMDS, 90k Intracranial injuries were predicted by Xgboost algorithm compared to 15k based on the primary diagnoses in the TMDS. However these findings need to be confirmed my reviewing the clinical notes.

Table 1. Predicting the ICD-9 CM Categories on all the TMDS data

ICD-9 subcategory	Frequency	Percent
Sprain	134608	43.20%
Intracranial	92590	29.70%
Injury other head	27716	8.90%
Open wound lower limb	10719	3.40%
Open wound upper limb	10707	3.40%
Fracture lower limb	7166	2.30%
Fracture upper limb	6675	2.10%
Open wound head neck trunk	5237	1.70%
Burns	4964	1.60%
Dislocation	4688	1.50%
Internal injury	1404	0.50%
Superficial	1198	0.40%
Contusion	941	0.30%
Amputation lower limb	924	0.30%
Nerve injury	665	0.20%
Fracture spine trunk	496	0.20%
Fracture face skull	494	0.20%
Injury blood vessel	66	0.00%
Amputation upper limb	48	0.00%
Total	311306	100.00%

- **What opportunities for training and professional development has the project provided?**
  - "Nothing to Report."
- **How were the results disseminated to communities of interest?**

  - Preliminary results were presented orally at the Navy Applications of Machine Learning (NAML) Conference in March 2021. NAML is an annual event showcasing current machine learning research relevant to naval applications. Participants include individuals from the defense community, industry, and academia.

- **What do you plan to do during the next reporting period to accomplish the goals?**

  - The accuracy of the ICD-9-CM data in the TMDS has never been quantified, largely because most of the data has been captured in Iraq and Afghanistan without any medical billing concerns. Sensitivity, positive predictive rates and accuracy metrics will be compared for Specific Aims 1-3.

4. **IMPACT:**

- **What was the impact on the development of the principal discipline(s) of the project?**

  - This research developed a set of programming rules or machine learning (ML) algorithms (i.e., algorithms generated by automated learning from manually coded data) to determine the principal ICD-9-CM code, which is the diagnosis code that is primarily responsible for the admission of the patient to the hospital. Using machine learning algorithms and NLP to develop a dataset through automating or semi-automating a coding system for deployed military personnel EHRs in an operational setting is a game changer. This research fills a void and reduces a heavily human capital project that is prone to subjective results and errors. Additionally, this effort proves that the military can take previously uncoded encounters and utilize the clinical notes to bring structure to unstructured data.

- **What was the impact on other disciplines?**

  - "Nothing to Report."

- **What was the impact on technology transfer?**

  - "Nothing to Report."

- **What was the impact on society beyond science and technology?**

  - "Nothing to Report."

5. **CHANGES/PROBLEMS:**

- "Nothing to Report"

6. **PRODUCTS:**

- **Publications, conference papers, and presentations**

  - **Journal publications.**

    - "Nothing to Report"

  - **Books or other non-periodical, one-time publications.**

    - "Nothing to Report"

  - **Other publications, conference papers, and presentations.**

    - Zouris, J. M., MacGregor, A. M., Olson, A. S., D'Souza, E., & Elkins, T. (2021, March 23-25). Medical Coding the Clinical Text Data from the Theater Medical Data Store

Using Natural Language Processing [Conference presentation]. Naval Applications of Machine Learning Workshop, Virtual event.

- **Website(s) or other Internet site(s)**
  "Nothing to Report"

- **Technologies or techniques**
  "Nothing to Report"

- **Inventions, patent applications, and/or licenses**
  "Nothing to Report"

- **Other Products**

  - We have begun to develop the Automate Clinical Repository (ACER) as part of Aim 5. This database will identify and predict the coding for trauma patients present in the TMDS that have not been coded in the EMED which will enable trauma patients to be expedited into the registry with considerably less clinical review.

7. **PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS**

- **What individuals have worked on the project?**

Name:	*LCDR Andrew Olson*
Project Role:	*Principal Investigator*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*2*
Contribution to Project:	*Project Management and Accountability*
Funding Support:	*Military*

Name:	*Edwin D'Souza*
Project Role:	*Statistician*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*4*
Contribution to Project:	*Data cleaning, algorithm development, application*
Funding Support:	*Leidos contractor*

Name:	*Trevor Elkins*
Project Role:	*Statistician*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*4*
Contribution to Project:	*Data cleaning, Natural Language Processing*
Funding Support:	*Leidos contractor*

Name:	*James Zouris*
Project Role:	*Statistician*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*4*
Contribution to Project:	*Data cleaning, algorithm development and test, ACER development*
Funding Support:	*Govt Employee*

Name:	*Andrew MacGregor*
Project Role:	*Research Epidemiologist*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*1*
Contribution to Project:	*ACER Development and data application*
Funding Support:	*Axiom Contractor*

Name:	*Tracy Negus*
Project Role:	*Division Manager*
Researcher Identifier (e.g. ORCID ID):	*1234567*
Nearest person month worked:	*1*
Contribution to Project:	*Project Management and Accountability*
Funding Support:	*Govt Employee*

- **Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?**
  - "Nothing to Report."
- **What other organizations were involved as partners?**
  - "Nothing to Report."

8. **SPECIAL REPORTING REQUIREMENTS**
   - **COLLABORATIVE AWARDS:** "Nothing to Report."
   - **QUAD CHARTS:** "Nothing to Report."

9. **APPENDICES:** "Nothing to Report."