# Leveraging Clinical Concept Documentation in Notes for NLP-Driven Phenotyping of Gulf War Illness

Malihehsadat Chavooshi[1,2], Javad Razjouyan[1,2,3], Drew A Helmer[1,2,3], Sarah T Ahmed[1,2], Arash Maghsoudi[1,2]

**Affiliations:** [1]VA HSR&D Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, TX; [2]Department of Medicine, Baylor College of Medicine, Houston, TX; [3]Big Data Scientist Training Enhancement Program (BD-STEP), VA Office of Research and Development, Washington, DC;

**Contact:** Malihehsadat Chavooshi, Baylor College of Medicine, Houston, TX, **Email:** malihehsadat.chavooshi@va.gov

## BACKGROUND

- Gulf War Illness (GWI) is a chronic multi-symptom illness affecting 30% of 1990-1991 Gulf War veterans.
- GWI is often not accurately captured by structured data elements, such as diagnosis codes.
- NLP offers a powerful approach to extract meaningful, symptom-level information from unstructured notes.
- By combining NLP-extracted concepts with machine learning, we can identify and characterize GWI in real-world EHR data.

## OBJECTIVE

To develop an interpretable NLP model that uses clinical concept documentation from unstructured notes to classify patients as GWI-positive or GWI-negative.

## DATA & METHODS

**Data Source:** Veterans Health Administration (VHA) Electronic Medical Record (EMR) – Corporate Data Warehouse (CDW)

**Cohort Summary:** 4,967 clinical notes identified using **WRIISC note titles** in the VHA EMR and annotated for GWI status (3,696 GWI-negative; 1,271 GWI-positive)

**Concept Extraction:** MedCAT NLP extracted clinical concepts from unstructured notes (negation/affirmation not applied).

**Clinician Review:** Physicians removed non-informative and irrelevant concepts.

**Concept Grouping:** Related concepts were consolidated by clinical similarity into grouped and individual features.

**Statistical Filtering:** Concepts compared between GWI+ and GWI− groups; retained if p-value < 0.1

**Model Development:** Random forest (nested 5-fold cross-validation) identified the most predictive clinical concepts. The top 13 concepts were used to build an interpretable Decision Tree for GWI Classification.

**Binary Classification Task:** Models classified notes as GWI-positive or GWI-negative.

**Gold standard:** physician-annotated GWI status

**Evaluation Metrics:** Accuracy, Sensitivity, Specificity, F1-score, ROC-AUC

## STUDY WORKFLOW

1. **Clinical WRIISC Notes** — Unstructured clinical notes
2. **MedCAT NLP Extraction** — 186 concepts
3. **Physician Review** — 152 concepts (34 removed)
4. **Clinical Grouping** — 53 features (27 grouped + 26 individual)
5. **Statistical Filtering** — 45 significant features ($p \leq 0.1$)
6. **Random Forest** — Feature ranking and importance
7. **Top 13 Concepts** — ~93% cumulative Random Forest feature importance
8. **Decision Tree** — Interpretable model built on top 13 concepts
9. **Prediction** — GWI+ or GWI−

## MODEL APPROACH: RANDOM FOREST + DECISION TREE

**Random Forest (Feature Selection)**
- Ranked clinical concepts by predictive importance
- Highest predictive performance
- Used for feature selection

**Decision Tree (Interpretable Model)**
- Built using the top 13 concepts
- Generated transparent IF–THEN decision rules
- Designed for clinical interpretability

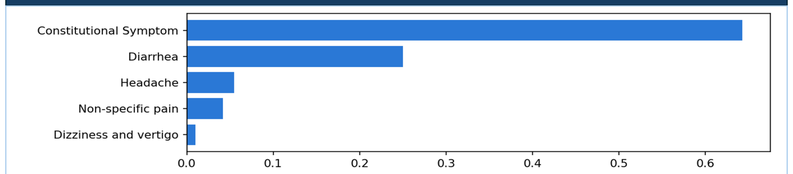
## RESULTS

| Metric | Random Forest Performance | Decision Tree Performance |
|---|---|---|
| Accuracy | **74.1%** | 71.9% |
| ROC-AUC | **0.769** | 0.737 |
| Sensitivity (GWI+) | 63.7% | 58.5% |
| Specificity (GWI−) | 77.7% | 76.5% |
| PPV (GWI+) | 49.6% | 46.2% |
| NPV (GWI−) | 86.2% | 84.3% |
| F1 Score | 0.557 | 0.516 |

## KEY FINDING

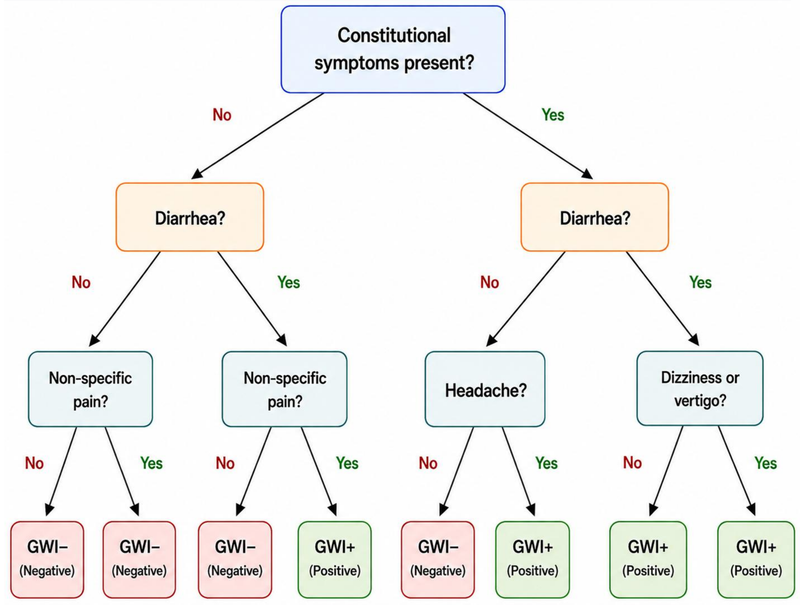
Random Forest achieved the highest predictive performance, whereas the Decision Tree generated transparent clinical decision rules using only five of the 13 concepts selected by the Random Forest.

## Top Predictive Clinical Concepts




The top **13 concepts** were selected based on **Random Forest** feature importance. The interpretable **Decision Tree** model retained **5** of these concepts.

## INTERPRETABLE DECISION TREE




## CONCLUSIONS

- NLP-derived clinical concepts produced moderately accurate discrimination between GWI-positive and GWI-negative clinical notes.
- Random Forest achieved the highest predictive performance for GWI note classification.
- The Decision Tree generated transparent clinical decision rules using only five of the 13 concepts selected by the Random Forest.

## FUTURE WORK

- External validation in independent cohorts
- Incorporate negation/affirmation detection
- Expand the concept set and perform longitudinal evaluation
- Evaluate clinical utility for decision support

## FUNDING

Veterans Health Administration Health Outcomes Military Exposures (HOME) service (formerly (Post-Deployment Health Services) and the Center for Innovations in Quality, Effectiveness and Safety (VA HSRD CIN 13-413) at the Michael E. DeBakey VA Medical Center, Houston, Texas.

## REFERENCE

Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*. 2021;117:102083.