Clinical Valence Testing

Lead/Mentor: *Faculty Lead*

Contributors: *Student Contributors*

Current Funding:

Future Funding:

IRB #:

RMID:

SPARCRequest:

Project Summary

This project implements behavioral testing of clinical NLP models to measure how valence-laden language (pejorative, laudatory, neutral descriptors) affects ICD diagnosis predictions. Building upon van Aken et al. (2021), the framework systematically applies linguistic perturbations to clinical texts and measures resulting shifts in model outputs.

Research Questions

1. How do pejorative patient descriptors (e.g., "non-compliant," "drug-seeking") affect ICD diagnosis predictions?

2. How do laudatory patient descriptors (e.g., "compliant," "cooperative") influence model outputs compared to neutral baselines?

Methods

Data Description

Dataset: MIMIC-III Clinical Notes

Format: CSV with clinical text and ICD-9-CM diagnosis codes

- `text`: Clinical note text (discharge summaries)
- `short_codes`: 3-digit ICD-9-CM codes
- 1,266 unique diagnosis codes after frequency filtering (min 100 occurrences)

Files:

- `DIA_GROUPS_3_DIGITS_adm_test.csv` - Test dataset
- `ALL_3_DIGIT_DIA_CODES.txt` - Reference diagnosis codes

Model Description

Model: DATEXIS/CORe-clinical-diagnosis-prediction

- Architecture: BERT-based transformer for multi-label ICD classification
- Pre-trained on MIMIC-III discharge summaries
- Output: Probability distribution over 1,266 ICD-9-CM codes

Configuration:

- Max sequence length: 512 tokens
- Batch size: 768 (H100 NVL GPU)
- Attention extraction: Layer 11, Head 11

Evaluation Approach

Valence Shift Types:

Shift	Description	Example Terms
Neutralize	Removes all valence terms	(baseline)
Pejorative	Negative descriptors	non-compliant, drug-seeking, difficult
Laudatory	Positive descriptors	compliant, cooperative, pleasant
Neutral	Objective descriptors	typical, presenting, evaluated

Statistical Tests:

- Paired t-test
- Wilcoxon signed-rank test
- Permutation test (Yeh 2000)
- Effect size: Cohen's d with FDR correction

Results

Output Files:

File	Content
<font name="Courier" size="9">{shift}_shift_diagnosis.csv</font>	Per-sample diagnosis probabilities
<font name="Courier" size="9">{shift}_shift_attention.csv</font>	Token-level attention weights
<font name="Courier" size="9">statistical_analysis.txt</font>	Statistical test results

Metrics:

- Mean probability shift per diagnosis code
- Number of significantly affected diagnoses ($p < 0.05$, FDR-corrected)
- Effect sizes (Cohen's d)

Resources

Code Repository: `clinical-valence-testing/`

Entry Point:

```
python main.py \
--test_set_path ./data/DIA_GROUPS_3_DIGITS_adm_test.csv \
--model_path DATEXIS/CORe-clinical-diagnosis-prediction \
--shift_keys neutralize,pejorative,laud,neutralval
```

Dataset: MIMIC-III (requires PhysioNet credentialed access)

Acknowledgements

Lead/Mentor: *Faculty Lead*

Contributors: *Student Contributors*

Past Contributors: *Student Contributors*

Current Funding:

Future Funding:

References

van Aken, B., Herrmann, S., & Loser, A. (2021). What Do You See in this Patient? Behavioral Testing of Clinical NLP Models. *Bridging the Gap: From Machine Learning Research to Clinical Practice, Research2Clinics Workshop @ NeurIPS 2021*.

van Aken, B. et al. (2021). Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. *EACL 2021*.