

Lab Valdes Research Traineeship



Summer 2025 – The Blue Scrubs & Clinical ML Projects

Talia Berler

August 7, 2025

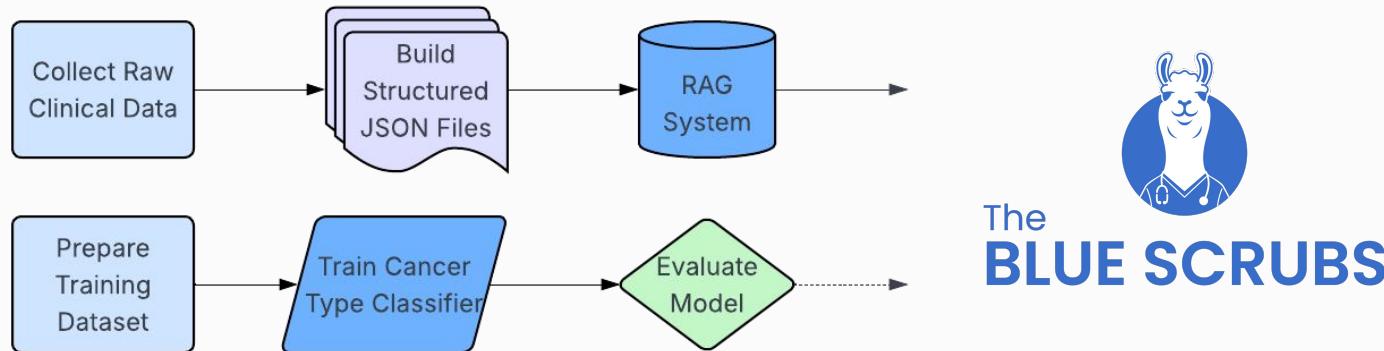


Traineeship Overview

Completed an 8 week internship with Lab Valdes performing data wrangling and preparing ML pipelines for The Blue Scrubs

Focus: Clinical dataset curation & machine learning pipeline design

Mission: Advance The Blue Scrubs—an open-source clinical AI platform



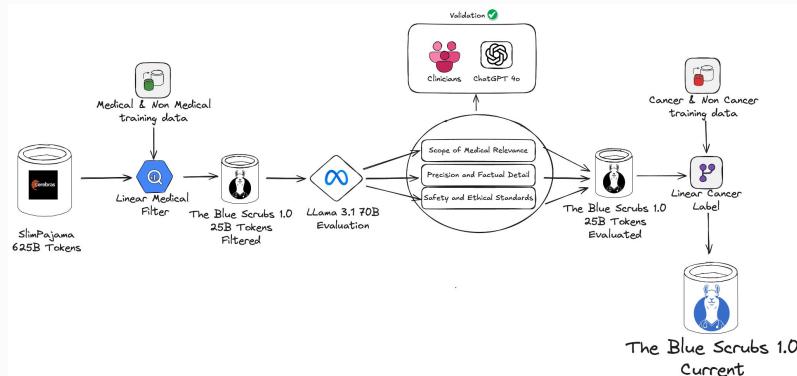
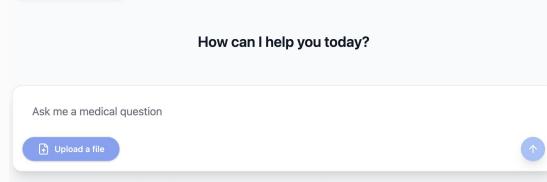
Key Responsibilities

1. Clinical guideline data for Virtual Clinician RAG system

1.1. Specialized clinical LLM-powered agents

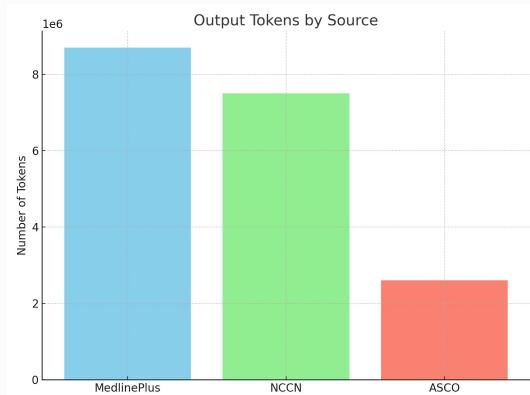
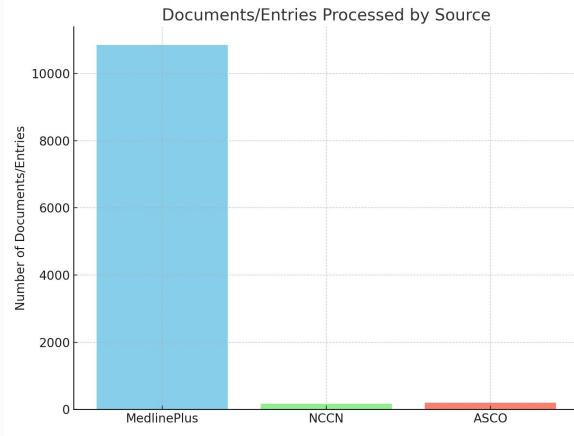
2. Build cancer type classifier for The Blue Scrubs dataset v2.0

2.1.1. 692B tokens of medical data, 10x larger than all existing medical text datasets



Data Collection I: Clinical Guidelines

	Documents	Tokens	Approach	JSON Fields
NCCN	161	7.5M	PDF parsing (PyMuPDF)	Category: (str) Citation: (str) Link: (str) Corpus: (str)
ASCO	203	2.6M	PDF parsing (PyMuPDF) NCBI API → PubMed citations	Title: (Str) Citation: (str) Link: (str) Corpus: (str)
MedlinePlus	10,854	8.7M	Web scraping (BeautifulSoup) Bulk XML parsing (lxml)	Meta: (obj) <ul style="list-style-type: none"> - Title: (str) - Link: (str) - Updated Date: (str) - References: (list of strings) - Aliases: (list of strings) - Related Topics: (obj: keys → topic type, values → obj: keys → topic title, values → URL) Text: (str)



Data Collection II: Clinical Guidelines Results



MedlinePlus

ASCO

MedlinePlus > Data > Cleaned Data > { } medline Encyclopedia.json > { } 0 >  Text

```
1 [ { 2
3   "Meta": {
4     "Title": "A guide to clinical trials for cancer",
5     "Link": "https://medlineplus.gov/ency/patientinstructions/000823.htm",
6     "Citation": "A.D.A.M. Medical Encyclopedia [Internet]. Johns Creek (GA): Ebix, Inc., A.D.A.M. Inc.; 2024. Available from: https://medlineplus.gov/ency/patientinstructions/000823.htm",
7     "Updated Date": null,
8     "Reviewed Date": "2024-08-21",
9     "References": [
10       "American Cancer Society website. Clinical trials. www.cancer.org/cancer/managing-cancer/other-cancer-treatments/clinical-trials.html",
11       "National Cancer Institute website. Clinical trials information for patients and caregivers. www.cancer.gov/cancer-information-for-patients/clinical-trials",
12       "National Institutes of Health website. Clinical Trials.gov. www.clinicaltrials.gov. Accessed 2024-08-21."
13     ],
14     "Aliases": [
15       [
16         "Intervention study - cancer"
17       ]
18     ],
19     "Related Topics": {
20       "Clinical Trials": "https://medlineplus.gov/clinicaltrials.html"
21     }
22   },
23   "Text": "A guide to clinical trials for cancer\n\nIf you have cancer (./article/001289.htm),"
24 },
25 ]
```

```
NCCN_Guidelines > {} NCCN_Guidelines.json > ...
1  [
2  {
3    "Category": "physician",
4    "Citation": "Referenced with permission from the NCCN Clinical Practice Guidelines in Oncology \n(NCCN Guidelines in Oncology, Version 2.2025, 01/17/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org).",
5    "Link": "https://www.nccn.org/professionals/physician_gls/pdf/ped_cns.pdf",
6    "Corpus": "Version 2.2025, 01/17/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org",
7  },
8  {
9    "Category": "physician",
10   "Citation": "Referenced with permission from the NCCN Clinical Practice Guidelines in Oncology \n(NCCN Guidelines in Oncology, Version 2.2025, 05/12/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org).",
11   "Link": "https://www.nccn.org/professionals/physician_gls/pdf/antiemesis.pdf",
12   "Corpus": "Version 2.2025, 05/12/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org",
13 },
14 {
15   "Category": "physician",
16   "Citation": "Referenced with permission from the NCCN Clinical Practice Guidelines in Oncology \n(NCCN Guidelines in Oncology, Version 1.2025, 05/09/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org).",
17   "Link": "https://www.nccn.org/professionals/physician_gls/pdf/smoking.pdf",
18   "Corpus": "Version 1.2025, 05/09/25 2025 National Comprehensive Cancer Network (NCCN), All rights reserved. NCCN.org",
19 },
20 }
```

```
1 ASCO_Guidelines > {} asco_guidelines.json > {} 0 > Corpus
2 [
3   {
4     "Title": "Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline.",
5     "Citation": "Vogelbaum MA, Brown PD, Messersmith H, Brastianos PK, Burri S, Cahill D, Dunn
6     Mohile N, Redjal N, Stevens G, Sulman E, van den Bent M, Wallace HJ, Weinberg JS, Zadeh G; Journal of clinical oncology : official journal of the American Society of Clinical Oncology 34932393.",
7     "Link": "https://doi.org/10.1200/JCO.21.02314",
8     "Corpus": "ASCO special articles Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline M
9     Messersmith, MPH; Priscilla K. Brastianos, MD; Stuart Burri, MD; Dan Cahill, MD, PhD; I
10    MD, PhD; Vinai Gondi, MD; Justin T. Jordan, MD; Andrew B. Lassman, MD; Julia Maues
11    PhD; Erik Sulman, MD, PhD; Martin van den Bent, MD; H. James Wallace, MD; Jeffrey S
12    abstract PURPOSE To provide guidance to clinicians regarding therapy for patients with brain
13    Panel and conducted a systematic review of the literature. RESULTS Thirty-two randomized trials
14    the primary evidentiary base. RECOMMENDATIONS Surgery is a reasonable option for patients who
15    are more likely to benefit than those with multiple brain metastases and/or uncontrolled systemic
16    receive local therapy regardless of the systemic therapy used. For patients with asymptomatic
17    deferral is specifically recommended in this guideline. The decision to defer local therapy should
18    benefits and harms that the patient may experience. Several regimens were recommended for non-
19    with asymptomatic brain metastases and no systemic therapy options, stereotactic radiosurgery
20    unresected brain metastases, excluding small-cell lung carcinoma. SRS alone or the surgical
21    brain metastases. SRS, whole brain radiation therapy, or their combination are reasonable options
22    should be offered to patients who receive whole brain
23    . . . . .
```

Data Collection III: Clinical Trials

	Documents	Approach	JSON Fields (Search)	JSON Fields (Source)
Clinicaltrial s.gov	545,616	Field extraction from individual clinical trial JSONs	Meta: (obj) <ul style="list-style-type: none">- Nct_id: (str; National Clinical Trial #)- Title: (str)- Url: (str)- Status: (str)- firstPosted: (str)- lastUpdated: (str) Text: (str; Synopsis)	Meta: (obj) <ul style="list-style-type: none">- Nct_id: (str; National Clinical Trial #)- Title: (str)- Url: (str)- Status: (str)- firstPosted: (str)- lastUpdated: (str) Text: (str; Eligibility & Design, Operational Info, Synopsis, Detailed Description)

Data Collection IV: Clinical Trials Results

SEARCH (concise)

nct_id: NCT00071435
title: Brain Function in Primary Lateral Sclerosis
url: <https://clinicaltrials.gov/study/NCT00071435>
status: COMPLETED
firstPosted: 2003-10-23
lastUpdated: 2008-03-04

TrialID: NCT00071435
Title : Brain Function in Primary Lateral Sclerosis
==== Synopsis ====
This study will examine whether the motor cortex (the part of the brain that controls movement) works properly in patients with Primary Lateral Sclerosis (PLS).
Healthy volunteers between 40 and 75 years of age and patients with ascending PLS (a subset of PLS) may be eligible for this study.
Participants perform a finger-tapping reaction time exercise while brain wave activity (electroencephalography, or EEG) is recorded.
For the EEG, brain activity is recorded by placing electrodes (small metal discs) on the scalp with an electrode cap.
Participants also undergo magnetic resonance imaging (MRI). This test uses a strong magnetic field and radio waves to create detailed pictures of the brain.
Conditions:
Motor Neuron Disease
Keywords:
EEG; Voluntary Movement; Motor Cortex; Reaction Time; Corticospinal; Primary Lateral Sclerosis; PLS; Healthy Volum

search.json

- Fewer token descriptions
- Indexable search file

SOURCES (clean b1+b2+b3+b4)

nct_id: NCT00071435
title: Brain Function in Primary Lateral Sclerosis
url: <https://clinicaltrials.gov/study/NCT00071435>
status: COMPLETED
firstPosted: 2003-10-23
lastUpdated: 2008-03-04

TrialID: NCT00071435
Title : Brain Function in Primary Lateral Sclerosis
==== Eligibility & Design ====
StudyType : OBSERVATIONAL
Phases : N/A
PrimaryPurpose : N/A
InterventionModel: N/A
Masking : N/A
SamplingMethod : N/A
BiospecRetention : N/A

Sex : ALL
AgeMin : N/A
AgeMax : N/A
Healthy : True

Eligibility Criteria:
EXCLUSION CRITERIA:

source.json

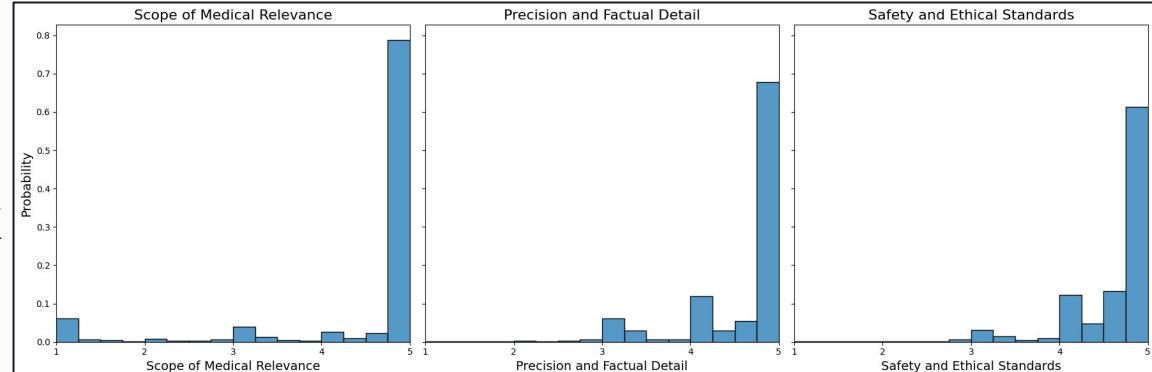
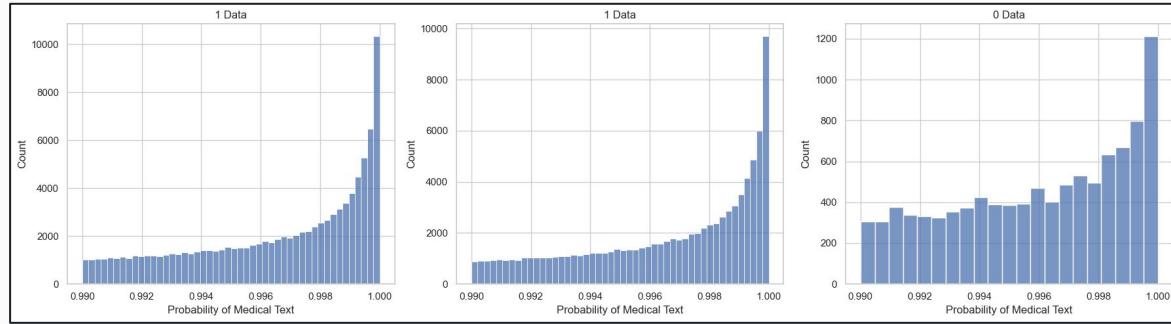
- Larger, in-depth descriptions for Virtual Clinicians

Classifier I: Training

Objective: Automate cancer-type labeling in Blue Scrubs v2.0 texts (proof-of-concept: breast cancer)

Sampling:

- Randomly drew 100K docs from 11.5 M TBS v1.0
 - a. $\geq 99\%$ medical probability
 - b. > 3 relevance, precision, and safety scores
- 90K cancer-flagged texts + 10K non-cancer for keyword QA



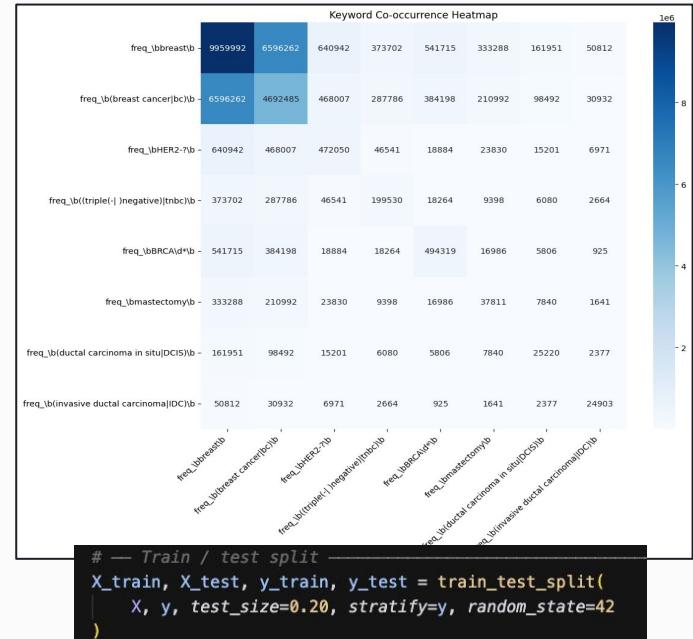
Classifier II: Training

Keyword Bootstrapping:

- Final regexes: `breast cancer`, `HER2-?`,
`(triple(-|)negative|tnbc)`, `BRCA\d*`
- 15 K positive & 15 K negative breast cancer samples

Pipeline:

- TF-IDF (uni/bi-grams, L2 normalization)
→ Logistic Regression
(`sklearn.linear_model`)
- 80/20 train/test split
- Runtime < 3 min CPU



```
# — TF-IDF feature extraction —
tfidf = TfidfVectorizer(
    max_features=10_000,
    ngram_range=(1,2),
    stop_words="english",
    lowercase=True
)
```

```
# — Logistic Regression classifier —
clf = LogisticRegression(
    max_iter=1_000,
    n_jobs=-1,
    class_weight="balanced",
    verbose=1
)
```

Classifier II: Model Evaluation

Data-leak risk:

- Training subset comes from same master corpus as deployment target
- Uses **0.26 %** of TBS v1.0 → **Exposure minimal**
 - Projected < 0.1 % of TBS v2.0

Results (hold-out):

```
Accuracy : 0.8635
Precision: 0.9014
Recall   : 0.8163
F1 Score : 0.8567
ROC-AUC   : 0.9446
```

Results (out-of-sample):

```
FLAG Out-of-sample metrics (ground truth = all positives):
Accuracy : 0.8209
Precision: 1.0000
Recall   : 0.8209
F1 Score : 0.9016
```

Future Directions

- **Staying on as Research Trainee** → continuing guideline extraction for TBS virtual clinicians
- **Application of classifier to TBS v2.0** → provide detailed cancer type labels beyond breast cancer on massive medical text dataset



The image shows a screenshot of a medical reasoning interface. At the top, there are three tabs: 'Answer', 'Sources (22)', and 'Medical Reasoning (10)'. The 'Medical Reasoning (10)' tab is currently selected, indicated by a blue underline. Below the tabs, there are two main sections: 'Step 1: Search in NCCN database' and 'Step 2: Search in ASCO database'. The 'Step 1' section contains a question and an 'Answer' section. The question is: 'What are the current NCCN guideline-recommended first-line treatment regimens for double-hit diffuse large B-cell lymphoma (DLBCL) with MYC and BCL2 rearrangements in a woman in her early 50s with Ann Arbor stage III disease and intermediate-high risk (IPI=3)?' The 'Answer' section provides a detailed response: 'For stage III diffuse large B-cell lymphoma (DLBCL), the NCCN guidelines recommend anthracycline-based chemoimmunotherapy regimens as first-line therapy. Preferred regimens include RCHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone) and Pola-R-CHP (polatuzumab vedotin-piiq, rituximab, cyclophosphamide, doxorubicin, prednisone). Dose-adjusted EPOCH (etoposide, prednisone, vincristine, cyclophosphamide, doxorubicin) plus rituximab is also listed as a recommended regimen, particularly for certain types of DLBCL. The guidelines specifically note that these regimens are advised for patients with stage III/IV DLBCL, which includes those with double-hit features such as MYC and BCL2 rearrangements.' Below the 'Answer' section, there is a 'Sources' link with a dropdown menu showing '2 sources'. The 'Step 2' section is partially visible at the bottom.



Thank you!

For further questions & comments:

talia.berler@moffitt.org | tkb40@miami.edu

—

Talia Berler
M.S. Data Science '26
University of Miami