Title: Development of the Open-Source Clinical Trial Selector for the Department of Veterans Affairs: Application of Natural Language Processing for Matching Inclusion Criteria

Authors: Nikhil Pesaladinne(1), Srithan Ram Thammineni(1), Ethan Ocasio(1), Riya Tadi(1) ..., Rafael Fricks(2), Gil Alterovitz(2)

(1): CTS group, GirlsComputingLeague

(2): National Artificial Intelligence Institute, Department of Veterans Affairs

Corresponding Author: Nikhil Pesaladinne; Washington DC VAMC Research Service, (151)
50 Irving Street, Washington, DC, 20422; 7039359573; nikhilpesala@gmail.com

ABSTRACT:

Purpose: Despite the rapid acceleration of clinical research progression as seen by the volume of clinical trial results in recent years, clinical trial research still faces barriers due to low rates of trial participation. One of those barriers specifically is the ability for patients to find applicable clinical trials. We propose the Clinical Trial Selector (CTS) as a solution to bridge the gap between patients and eligible clinical trials.

Methods: CTS is a web-based application designed to match veteran patients to clinical trials where they may be eligible. It uses the VA Lighthouse and CMS BlueButton application programing interfaces (APIs) to access patient electronic health record (EHR) data and match it with NCI and NIH clinical trial registries. CTS's automation simplifies the process of searching for trials. The application's advanced filtering process increases its feasibility and differentiates it from other possible solutions.

Results: The results demonstrate that there is no significant statistical correlation between common cancers and the number of clinical trials, as well as the top one hundred principal diagnoses in the U.S community hospitals and their respective number of clinical trials available, with a p-value less than 0.01. Some medical conditions were also statistical outliers, suggesting that there is a need to bridge the gap between conditions that are over/under-studied. Results from both tests have shown that applications like CTS are necessary when connecting patients to respective clinical trials. Applications such as these could increase the rate at which life-saving treatments are discovered.

Conclusion: There is a strong need for applications like CTS to connect patients to eligible clinical trials, due to the weak correlation between the number of clinical trials and prevalence of cancers.

INTRODUCTION:

Clinical trials frequently struggle to recruit enough eligible patients. One identified barrier is that patients often are not aware of clinical trials that may be suitable in their case.[15] As a result, fewer patients participate, contributing to patient recruitment delays that become a rate-limiting step in the discovery of potentially life-saving therapies and drugs. The widespread use of electronic health records (EHR) in the U.S. over the past decade has allowed for the use of automated, electronic screening of trials for patients which direct the patient to a select few more relevant clinical trials.[10,14] However, many existing trial finders still require manual information input systems that give less precise and less accurate clinical trials as results to users. The need for manual entry either places additional burden on the provider, who must search on behalf of their patients, or on patients and their families, who must navigate nuanced language . Thus, there is no straightforward way to automate the matching of eligible patients to trials. Many of the current clinical-trial search engines, such as the NIH's clinical trial search engine, offer large numbers of clinical trials to filter through.

BACKGROUND ON CLINICAL TRIAL MATCHING:

The results were prepared using a Pearson's correlation coefficient test. We compared the thirty-four most common cancers among men and women worldwide with their respective number of clinical trials available. The list of cancers was retrieved by the GLOBOCAN 2018 estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer1. Their corresponding number of clinical trials was retrieved using the National Cancer Institute's (NCI) Cancer Clinical Trials Search API.

The list of the top 100 principal diagnoses in the U.S community hospitals, their respective number of patients diagnosed, and their respective mortality rates were retrieved from the Nationwide Inpatient Sample of the Healthcare Cost and Utilization Project (HCUP)[4]. In addition, the corresponding number of available clinical trials was recovered from the clincicaltrials.gov database.

Figure 1 displays a scatter plot with each condition and its respective number of cases and available clinical trials. Table 1 and Table 2. display the data used to create this scatterplot. Figure 2 compares the relative prevalence of disease to the number of associated clinical trials. A larger circle indicates a more widespread disease, with a darker shade of blue indicating a greater number of clinical trials.

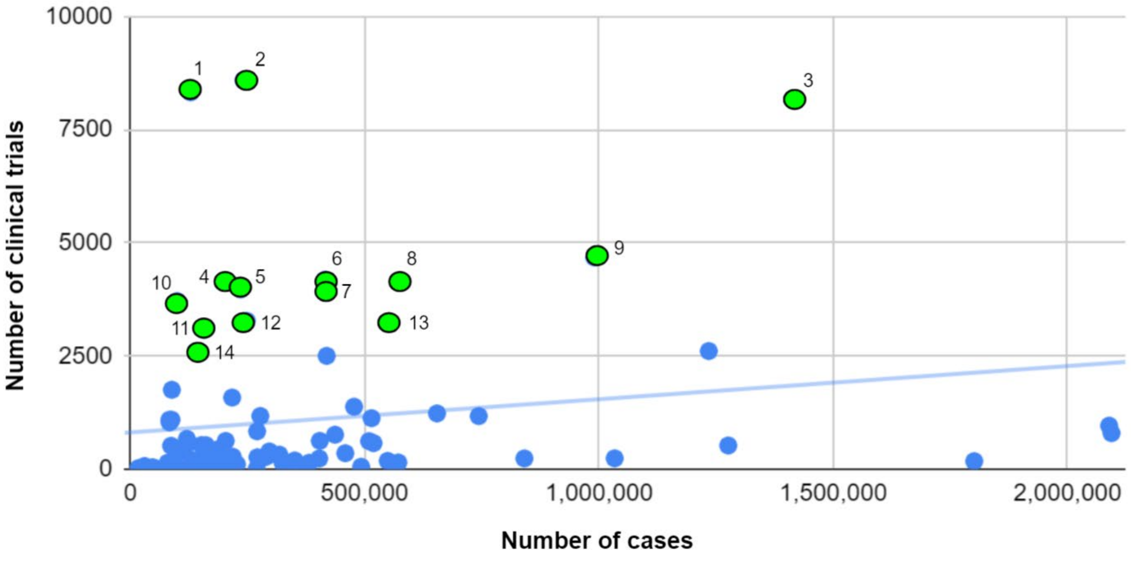


Figure 1. Scatterplot between number of cases and number of trials with outliers highlighted and line of best fit. (1) HIV Infections, (2) Hypertension, (3) Coronary atherosclerosis, (4) Pancreatic disorders, (5) Substance-related mental disorders, (6) Osteoarthritis, (7) Asthma, (8) Cardiac dysrhythmias, (9) Congestive heart failure, nonhypertensive, (10) Deficiency and other anemia, (11) Esophageal disorders, (12) Schizophrenia, (13) Chronic obstructive pulmonary disease and bronchiectasis, (14) Peripheral and visceral atherosclerosis

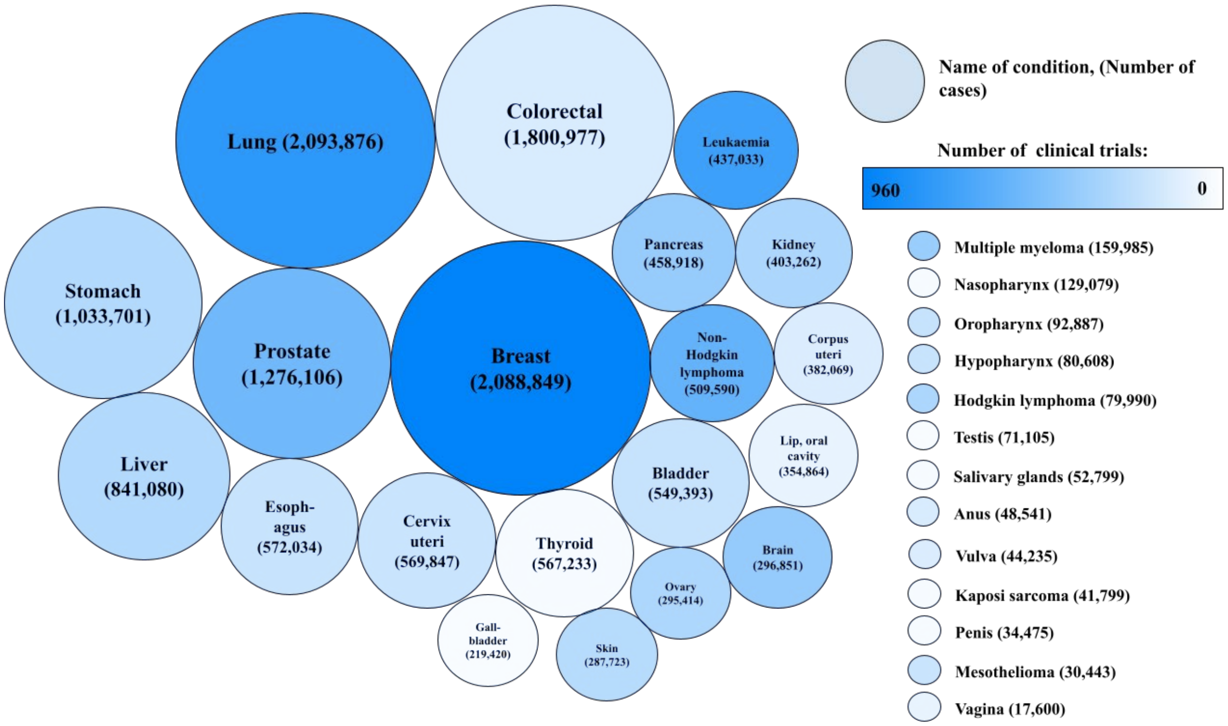


Figure 2. Proportionality diagram comparing number of clinical trials with number of cases

Although there seems to be a general positive correlation between the number of cases for particular cancer and the number of trials, after performing the Pearson's correlation coefficient test, we notice that there is not a significant correlation between the data (r2 = 0.4812, p = 0.0039 < 0.01). This suggests that it is not necessarily true that a common condition has a proportional number of clinical trials (e.g., Thyroid Cancer). The opposite is also not true; less common conditions might have an excess of clinical trials (e.g., Leukemia). This suggests that difficulty could arise where a majority of a population with a condition struggles to find a relevant clinical trial or investigators themselves struggle to find patients with the condition they require.

A similar test was done with U.S Community hospitals' conditions and their respective number of available clinical trials. However, there were some outliers and arbitrary values in the data. Diagnoses with no corresponding conditions and diagnoses that covered a large category of conditions were excluded, there were marked by "*" and "**" respectively. After this normalization, the data did not show a correlation and was significant ($r^2 = 0.163$, $p=.001503 < 0.01$). This data set also supports the previous discussion of possible difficulties that arise during the recruitment of clinical trials because the same trends arise compared to clinical trial data with cancer patient data.

An outlier test suggests some abnormalities (highlighted in green), with conditions having only a limited number of patients for a surplus number of available clinical trial spots. Specifically, Hypertension, HIV Infections, Coronary atherosclerosis, Pancreatic disorders, Congestive heart failure, Osteoarthritis, Substance-related mental disorders, Cardiac dysrhythmias, Asthma, Deficiency, and other Anemia, Schizophrenia, Esophageal disorders, Chronic obstructive pulmonary disease, and Peripheral and Visceral atherosclerosis. Some of these conditions are among the most common in the world (Coronary atherosclerosis, Congestive heart failure, Cardiac dysrhythmias), suggesting that programs such as CTS have apt applications in the real world for increasing the accessibility of trials for patients.

Existing clinical trial match tools

There are various tools in use right now for helping patients find clinical trials. In evaluating the strengths of these approaches, we identified two main limitations in existing tools: 1) the reliance on manual entry or questionnaires, 2) narrow focus in trial matching. We elaborate on these limitations in context.

To make their product available to the general population, most clinical trial finders apply a questionnaire to filter out a select few trials. The most common issue apparent so far is the reliance on the questionnaire system for filtering clinical trials. Although this system may be acceptable to use once or twice, it would become tedious for recurring users. Furthermore, the questionnaire system does not allow for a thorough and meticulous search because it depends on superficial information instead of real medical data.

- For example, the questionnaire in the ResearchMatch Trials Today website is not liable to differentiate between a search concerning “breast cancer” and a search concerning “HER2+ breast cancer”. Instead, in both searches, the clinical trials results are the same and now the user is forced to look through the available trials and manually filter them. ResearchMatch Trials Today is a clinical trials finder developed by Vanderbilt Institute of Clinical and Translational Research. This trials selector uses a questionnaire system to filter by trial locations, trial conditions (diseases), sponsors, and studies. Trials Today connects to unique qualifiers in the Unified Medical Language System (UMLS—curated by the NIH National Library of Medicine) when a patient types a term or a phrase, such as “leukemia,” to gather synonyms for accessing all relevant trial records.[12] Trials Today can also find trials for up to 2,602 medical conditions. One of their most prominent activities in the past few years was a nationwide marketing campaign to publicize their website that included mostly Facebook and other social media ads in targeted specific cities such as Miami or Washington DC.[5] One concern surrounding this application is an ethical one due to the usage of a tracking pixel. This pixel is only applied on the home page and not much deeper on the website, so it collects, according to the developers, information regarding the patients’ sex, age, and geographical location. This trial finder uses a questionnaire system to filter among trials in which users must enter personal information manually for the software to sift through the trial pool and match trials to the given information.

- DQueST is another trial finder based on a “dynamic questionnaire”. This trial finder works in separate parts. First, in the information extraction and curation part (after the patient enters search data), free-text eligibility criteria for 252,330 trials preserved in the ClinicalTrials.gov repository (as of August 2018) are analyzed and translated to a standardized format. Line breaks were used to divide the free-text eligibility criteria into paragraphs, which were subsequently separated into parts using a sentence splitting algorithm from Stanford Core NLP.[6,7] For domain prediction, the same model was employed. The negation status was then determined using NegEx[2], a rule-based method. One thing about DQueST that might be a concern to some is that DQueST is not yet easily accessible for the public to use as there does not seem to be a website accessible by Google for it. Secondly, as given in the description, DQueST still requires a questionnaire to filter. This does not differentiate it from almost any other clinic trial finder in that you must re-enter information every time it is used. Instead, if the application had security approved to safeguard and store private medical information, it would be much more convenient for the patient.

- The Janssen Global Trial Finder was made by the Janssen Pharmaceutical Company which is a part of Johnson & Johnson. The Janssen Global Trial Finder (JGTF) is another basic trial finder that is available to anyone through their website. It uses four queries to give results: first, a bar in which the user enters their zip code; then, a dropdown menu to select the condition for which the user wants to find a clinical trial; next, a bar for the user to select how far they are willing to travel; finally, a three-pronged filter system that allows the user to determine sex, age, and recruitment status of the clinical trials. The search then displays a list of clinical trials that fit the parameters set, along with an interactive Google Map that shows the address of the clinical trial, and links are provided to get more information regarding the selected clinical trials.

The other key limitation we identified is that many tools tend to have a highly specialized scope.

- The COVID-19 Trial Finder is a COVD-19 specific trial finder that uses a more demographically based questionnaire due to the varied effects of COVID-19 based on \ age-group and prior body condition. The initial questionnaire is only 5 questions long, which will produce a somewhat broad list of trial options available to the user. According to the developers, the COVID-19 Trials Finder sources its clinical trial options “from ClinicalTrials.gov by querying all the trials indexed with “COVID-19” being their condition.” Standard 5 question test leads to a 79.76% precision in matching trial locations for COVID-19.[13] CTS is quite different from this trial finder mainly because CTS finds trials for a myriad of patient conditions, including COVID-19, whereas this trials finder only focuses on COVID-19 (hence the name).
- The Fox Trial Finder, made by a section of Michael J. Fox Foundation, is similar to the COVID-19 trial finder in that it only finds trials for a small number of diseases. The foundation mainly focuses its efforts on Parkinson’s Disease but the trials finder finds clinical trials for Parkinson’s Disease and a few others. Namely: Multiple System Atrophy, Progressive Supranuclear Palsy, Lewy Body Dementia, and Corticobasal Degeneration. This is a disadvantage for the Fox trial finder because it only caters to a very small and specific population since it only provides results for 5 conditions.[9] This system also has a relatively small questionnaire on their website that is open for anyone to use without the need for login. By March of 2015, over 42,500 individuals had either made accounts or registered with the Fox Trial Finder.[3] Although there is potential for more accounts to be made with Fox Trial Finder,

> Moreover, there is no evidence that Fox Trial Finder uses a clinicaltrials.gov API or any such external API.

This is one place where CTS has an advantage; since CTS allows live medical data to be stored, it allows for a much deeper and more specific search that would indeed only display HER2+ breast cancer related results in the given situation. Furthermore, since CTS stores live medical data, there is no need for the user to fill out any questionnaires after login because a search can be made with all the stored information. This omits one major laborious step that all other trial-finders contain. Secondly, some of the other trial-finder in the field right now are designed to find trials for a few specific conditions. Examples of these include the COVID-19 Trial Finder and the Fox Trial Finder. By limiting their scope like this, such trial-finders are also limiting their reach to the public. This is another place where CTS has an advantage given that it accommodates searches for a myriad of conditions and has a potential reach of more than 50 million people in America. Lastly, CTS is authorized to handle sensitive and private medical information per HIPAA protocols, which is not something that can be said about any other trial-finder.

METHODS:

Application Description:

CTS is a web-based application that allows patients with records at Veteran Affairs (VA) and/or Medicare (CMS) to find clinical trials they are eligible for. Traditional clinical trial databases ask users to enter the relevant medical information directly as search criteria. CTS only requires users to log in through their government-issued accounts and using secure data authorization protocols; it retrieves demographic and clinical information, which it then uses to search for relevant trials in available databases, also applying natural language processing to obtain further filtering criteria out of the unstructured eligibility text descriptions.

The application is built with Python and runs on a Python webserver called Flask. It is an open-source application, and Anaconda is used as a package manager for all dependencies. The front-end was developed with HTML and JavaScript along with CSS Bootstrap following the stylistics of the Department of Veteran Affairs. All source code can be found on a GitHub repository, and code merges into production are managed through Git Hooks.

CTS runs on both development and production environments, and both are currently deployed on AWS EC2 instances; the live production site can be accessed here: https://cts.girlscomputingleague.org. Since patient data is sensitive and can pose a security risk, we do not have a persistent database. Instead, we cache all data in a session that a Redis ElastiCache handles. AWS also handles our application URL through Route 53, and we have an Amazon Load Balancer (ALB) set up across three zones in case of high traffic.

All data transfer and network requests to CTS are made through the Secure Shell (SSH) protocol, whose certificates are automatically generated by CertBot. User authentication is done through the authenticated pages provided by the VA and CMS, both of which utilize the OAuth 2.0 protocol (specifically OpenID Connect for VA) to securely provide authorization tokens to CTS once the patient authenticates. Once the patient completes the authentication process, they get redirected to the CTS app. Via OAuth 2.0 protocol, CTS seamlessly provides an access code, which it keeps in session and subsequently uses as an authorization token to request data to APIs on behalf of the patient. This token is only valid while the session is active (~ 1 hour).

To collect patient data and compare it with the eligibility criteria of a given clinical trial, CTS makes several API calls to public Rest APIs. Specifically for patient data itself, CTS uses the patient access code to make calls to the VA LightHouse and CMS BlueButton API, which access patient EHRs and make health information available through the API endpoints. However, the returned diagnosis information is in a different format than what the NCI API requires to search for clinical trials (SNOMED-CT and ICD-10 from VA and ICD-9 and ICD-10 from CMS). CTS invokes the Universal Medical Language System (UMLS) API to convert the given codes and the required NCIT codes. Lastly, to extract computer parsable eligibility criteria from the unstructured version, CTS implements the Facebook Clinical Trial Parser, built on the Go language. This returns a list of boolean statements which CTS compares with the given patient's health information and filters out the trials that the patient is not eligible for, displaying the finalized list on the website for the user to check available trials that they qualify for.

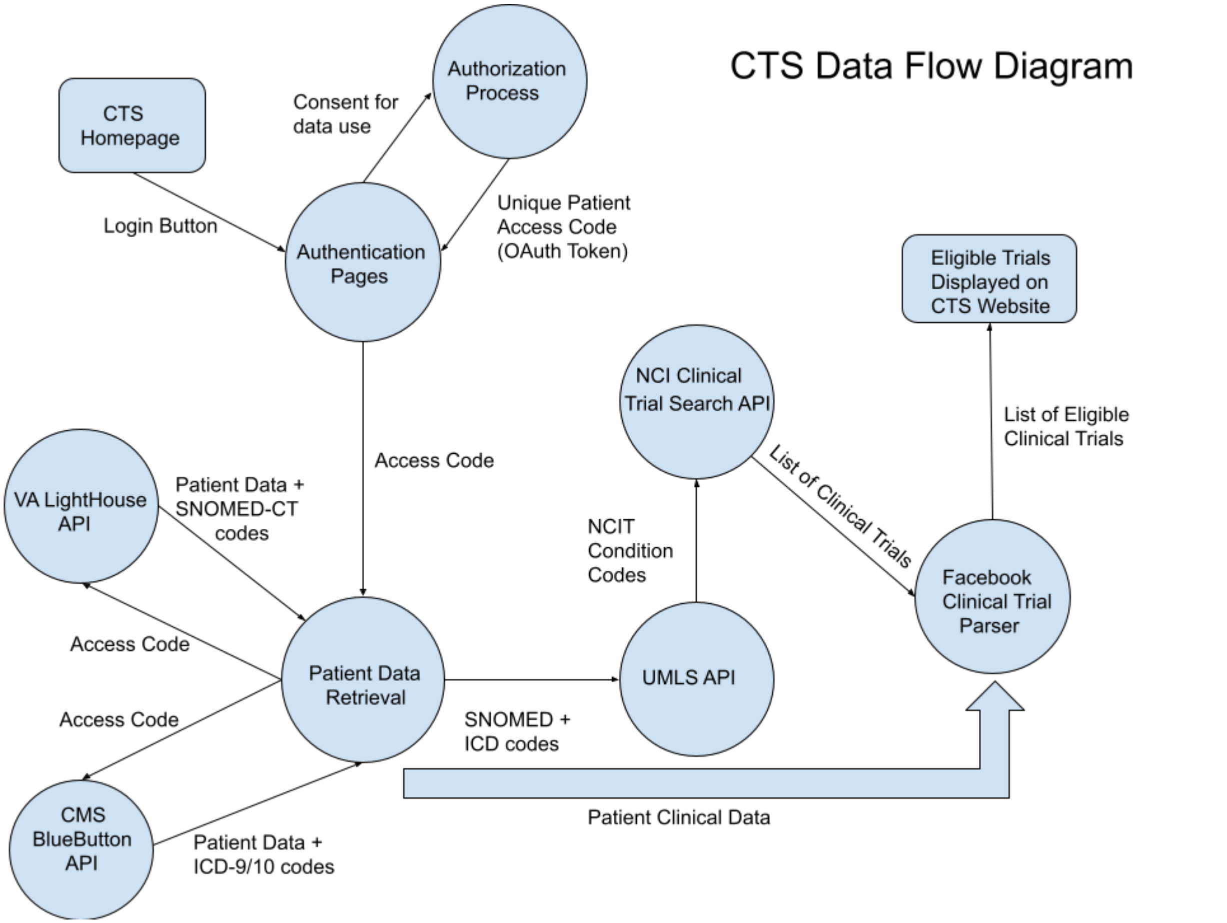
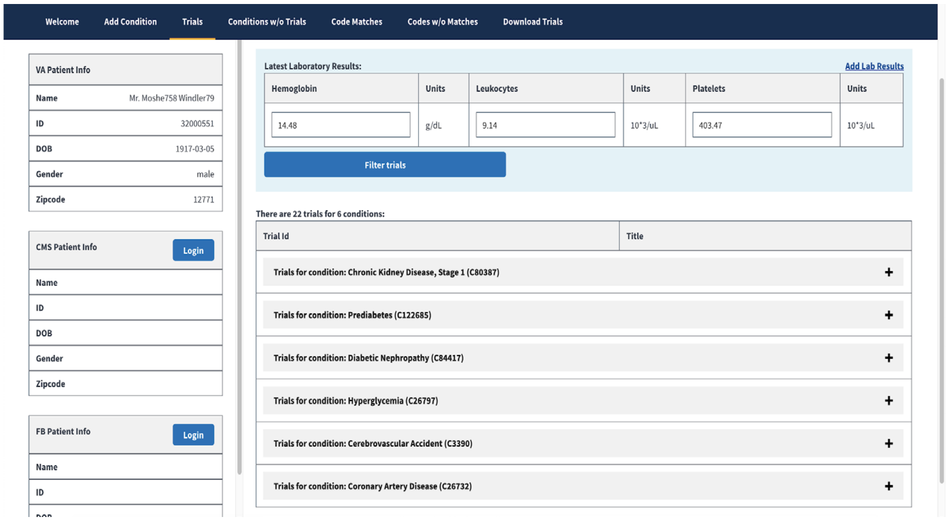


Figure 3 . Flowchart of CTS

RESULTS:

After logging in, the user is presented with the list of trials as seen in Figure 3 and lab results and other pre-populated biometric data based on the patient's records. Users may manually edit the results and conditions if they know any new changes that are not reflected in the database or application for querying trials, as CTS has read-only access to patient records. The CTS application enables users to search for a condition and then lists several related conditions using the UMLS similarity API to ensure the user picks the most specific condition to get the best results, differentiating the condition “Diabetes” and “Diabetes Insipidus” or “Type 1 Diabetes Mellitus” .



VA Patient Info	
Name	Mr. Moshe758 Windler79
ID	32000551
DOB	1917-03-05
Gender	male
Zipcode	12771

CMS Patient Info	
Name	
ID	
DOB	
Gender	
Zipcode	

FB Patient Info	
Name	
ID	

Latest Laboratory Results:

Hemoglobin	Units	Leukocytes	Units	Platelets	Units
14.48	g/dL	9.14	10^3/uL	403.47	10^3/uL

There are 22 trials for 6 conditions:

Trial Id	Title
Trials for condition: Chronic Kidney Disease, Stage 1 (C80387)	
Trials for condition: Prediabetes (C122685)	
Trials for condition: Diabetic Nephropathy (C84417)	
Trials for condition: Hyperglycemia (C26797)	
Trials for condition: Cerebrovascular Accident (C3390)	
Trials for condition: Coronary Artery Disease (C26732)	

Figure 4. Screenshot of CTS Homepage

The CTS tool also sorts trials by their condition, as seen above in Figure 4. Through this, users can only search through trials of a specific condition if that is what they’re interested in, despite having trials for other conditions they may have. Users can also see additional detail for each trial, including the eligibility criteria, description, and locations where the trial is located. The CTS application also lists conditions it could not parse or find trials for completeness. If a user finds the condition they are interested in in this list; the CTS tool indicates that it could not find any trials the user would be eligible for in the set.

Although users must log in to use the application, the application holds no patient records, causing the user to re-search for clinical trials every time they log out. This is by design: - to preserve security and ensure that no patient data will be leaked, the application only stores temporary session data, which is deleted when the patient logs out, effectively removing the data from CTS' records.

CTS differentiates itself from other general trial search tools through its filtering process. To start, it automatically filters trials that do not match the user's demographic information. Then, the user may add and modify the biological data and conditions list and then sift through their biological information. Next, the application utilizes the Context-Free Grammar (CFG) portion of the Facebook clinical trial parser to translate eligibility criteria into mathematical expressions using the inputted patient data. After the trials are filtered, the user can see which trials they are eligible for and the reason (for instance, the application will describe which eligibility criteria they failed and why).

DISCUSSION:

Clinical trials have clearly defined eligibility criteria that patients need to meet to provide meaningful results. Unfortunately, patients are often unable to locate clinical trials that fit them due to difficulty understanding nuances used to describe their condition precisely. Clinical Trial Selector is a novel approach that can take a patient's automated EMR data and match it to clinical trials eligibility to determine appropriate trials. CTS retrieves clinical trials from the National Cancer Institute's (NCI) Cancer Clinical Trials Search API and leverages natural language processing techniques to produce a list of eligible clinical trials for the patients using their clinical history. According to common cancer and available clinical trial statistics, results show a weak correlation between the number of cases and trials, further emphasizing the need for this application.

The graphs also suggest a consistent potential issue. Conditions that are well studied (i.e., those that have the most associated clinical trials) are almost always above the line of best fit, suggesting that those studies may potentially struggle to find eligible patients. There are multiple such conditions in both data sets (Non-Hodgkin lymphoma, Leukemia, Breast Cancer, Hypertension, HIV, etc..). Similarly, many conditions are understudied (i.e., have many more patients than associate clinical trials) that lie under the line of best fit, such as Colorectal Cancer and Pneumonia. These outliers highlight the importance of applications like the Clinical Trial Selector and potential cases where it would be of assistance.

Overall, the most common conditions stay moderately in shape with the line of best fit. Most are overstudied, meaning they have a relatively high number of clinical trials compared to their respective number of patients, but some are understudied, which poses a potential problem. These understudied conditions and outliers make up some of the most common in the world. The production of potentially life-saving medications is being slowed due to the lack of participation in clinical trials.

Appendix:

Table 1.

Type of Cancer	# of cases in 2018	# of trials (from NCI)
Lung	2,093,876	796
Breast	2,088,849	960
Colorectal	1,800,977	176
Prostate	1,276,106	523
Stomach	1,033,701	239
Liver	841,080	237
Oesophagus	572,034	147
Cervix uteri	569,847	131
Thyroid	567,233	47
Bladder	549,393	182
Non-Hodgkin lymphoma	509,590	619
Pancreas	458,918	352
Leukaemia	437,033	761
Kidney	403,262	235
Corpus uteri	382,069	142
Lip, oral cavity	354,864	102
Brain, central nervous system	296,851	391
Ovary	295,414	308

Melanoma of skin	287,723	265
Gallbladder	219,420	35
Larynx	177,422	90
Multiple myeloma	159,985	298
Nasopharynx	129,079	34
Oropharynx	92,887	118
Hypopharynx	80,608	72
Hodgkin lymphoma	79,990	142
Testis	71,105	10
Salivary glands	52,799	22
Anus	48,541	41
Vulva	44,235	31
Kaposi sarcoma	41,799	15
Penis	34,475	16
Mesothelioma	30,443	69
Vagina	17,600	23

Table 2.

2 - # of cases

3 - # of trials

4 – mortality rate of conditions

Liveborn *	3,827,230		0.32
Coronary atherosclerosis	1,417,661	8,149	0.95
Pneumonia	1,234,565	2,609	6.35
Congestive heart failure, nonhypertensive	990,085	4,679	5.16
Acute myocardial infarction	743,677	1,173	8.95
Trauma to perineum and vulva *	693,164		0
Acute cerebrovascular disease	654,600	1,231	10.57
Normal pregnancy and/or delivery *	578,841		0
Affective disorders *	574,120		0.1
Cardiac dysrhythmias	574,046	4,126	1.21
Chronic obstructive pulmonary disease and bronchiectasis	547,480	3,243	2.9
Spondylosis, intervertebral disc disorders	519,130	575	0.16
Chest pain	514,895	1,126	0.07
Fluid and electrolyte disorders	492,750	53	3
Biliary tract disease	477,660	1,379	0.74
Complication of device, implant or graft *	469,556		1.86

Fetal distress and abnormal forces of labor *	428,124		0.01
Septicemia	419,158	2,503	13.74
Asthma	418,227	3,964	0.42
Osteoarthritis	415,264	4,181	0.17
Urinary tract infection	404,458	625	1.66
Diabetes mellitus **	403,460		1.59
Other complications of birth, puerperium affecting management of mother *	379,223		0.03
Fracture of neck	351,033	191	3
Other complications of pregnancy *	341,656		0.04
Rehabilitation care, fitting of prostheses, and adjustment of devices *	335,978		0.97
Complications of surgical procedures or medical care *	335,365		1.63
Skin and subcutaneous tissue infection	325,223	136	0.58
Gastrointestinal hemorrhage	318,104	321	4.4
Alcohol-related mental disorders	277,610	1174	0.09
Intestinal obstruction	271,324	262	3.57
Fracture of lower limb	270,698	836	0.5

Early or threatened labor	269,756	8	0.01
Previous C-section *	268,295		0
Umbilical cord complication *	267,188		0
Secondary malignancies *	264,022		13.77
Maintenance chemotherapy, radiotherapy *	258,600		0.76
Schizophrenia	248,833	3,275	0.05
Hypertension	241,745	8,591	3.46
Substance-related mental disorders	235,490	3972	0
Diverticulosis and diverticulitis	227,673	105	1.4
Benign neoplasm of uterus	222,123	13	0.02
Appendicitis	218,668	273	0.21
Epilepsy	217,431	1,582	1.13
Polyhydramnios *	206,989	6	0
Acute bronchitis	203,700	623	0.2
Pancreatic disorders	198,187	4143	1.79
Transient cerebral ischemia	195,864	238	0.22
Syncope	189,193	182	0.27
Phlebitis, thrombophlebitis and thromboembolism *	188,566	3	1.15

Calculus of urinary tract	187,441	438	0.09
Hypertension complicating pregnancy, childbirth and the puerperium *	174,242		0.05
Aspiration pneumonitis	173,114	57	19.64
Occlusion or stenosis of precerebral arteries *	167,750		0.52
Intracranial injury	167,331	201	7.43
Other fractures *	166,785		1.33
Lower respiratory disease	161,760	174	3.08
Abdominal hernia	161,289	530	1.23
Cancer of Lung **	158,150		15.76
Esophageal disorders	158,065	3083	0.64
Prolapse of female genital organs	157,673	34	0.04
Malposition, malpresentation *	156,507		0.01
Gastrointestinal disorders **	155,470		2.45
Abdominal pain	152,026	532	0.42
Other and unspecified benign neoplasm *	151,628		0.59
Fetopelvic disproportion, obstruction *	149,209		0.01
Other mental conditions *	144,557		0.07
Gastritis and duodenitis	144,505	5	0.92

Fracture of upper limb *	142,830		0.43
Peripheral and visceral atherosclerosis *	140,667	2603	6.48
Senility and organic mental disorders	134,427	196	1.19
Noninfectious gastroenteritis *	130,516		0.16
HIV infection	128,760	8325	10.95
Cancer of breast **	125,663		1.5
Poisoning by other medications and drugs *	125,152		1.11
Intestinal infection	120,768	669	0.75
Hyperplasia of prostate	120,399	542	0.26
Cancer of colon **	118,489		5.48
Other female genital disorders *	110,649		0.12
Cancer of prostate **	107,054		1.61
Other nervous system disorders *	106,730		1.31
Forceps delivery	106,084	6	0
Other connective tissue disease *	104,078		0.63
Pleurisy, pneumothorax, pulmonary collapse	101,960	342	4.37
Viral infection	101,147		0.34
Prolonged pregnancy	101,001	32	0.01
Deficiency and other anemia	100,040	3712	1.97

Crushing injury or internal injury *	97,672		4.54
Heart valve disorders	88,783	1753	4.64
Other circulatory disease *	88,673		2.82
Acute and unspecified renal failure	88,120	1092	13.13
Endometriosis	87,128	517	0.02
Other bone disease and musculoskeletal deformities *	86,274		0.27
Sprains and strains *	86,184		0.09
Other upper respiratory infections *	85,288		0.11
Pulmonary heart disease	83,931	1034	6.77
Peri-, endo-, and myocarditis, cardiomyopathy	83,548	1090	5.81
Aortic, peripheral, and visceral artery aneurysms	83,317	106	13.15
Other injuries and conditions due to external causes *	81,793		2.08

* - Was not recognized as a condition

** - Excluded due to possible other locations of diagnosis

REFERENCES

1. Bray, Freddie, et al. "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, 12 Sept. 2018, pp. 394-424, https://doi.org/10.3322/caac.21492. Link
2. Chapman, Wendy W., et al. "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." *Journal of Biomedical Informatics*, vol. 34, no. 5, Oct. 2001, pp. 301-10, https://doi.org/10.1006/jbin.2001.1029. Link Google Scholar
3. Dorsey, E. Ray, et al. "Feasibility of Virtual Research Visits In Fox Trial Finder." *Journal of Parkinson's Disease*, vol. 5, no. 3, 2015, pp. 505-15, https://doi.org/10.3233/JPD-150549. Link Google Scholar
4. Elixhauser A, Steiner CA. Most Common Diagnoses and Procedures in U.S. Community Hospitals, 1996. Summary, HCUP Research Note. Agency for Health Care Policy and Research, Rockville, MD. http://www.hcup-us.ahrq.gov/reports/natstats/commdx/commdx.htm
5. Jerome, Rebecca N., et al. "To End Disease Tomorrow, Begin with Trials Today: Digital Strategies for Increased Awareness of a Clinical Trials Finder." *Journal of Clinical and Translational Science*, vol. 3, no. 4, Aug. 2019, pp. 190-98, https://doi.org/10.1017/cts.2019.404. Link Google Scholar
6. Liu, Cong, et al. "DQueST: Dynamic Questionnaire for Search of Clinical Trials." *Journal of the American Medical Informatics Association*, vol. 26, no. 11, 7 Aug. 2019, pp. 1333-43, https://doi.org/10.1093/jamia/ocz121. Link Google Scholar

7. Manning, Christopher, et al. "The Stanford CoreNLP Natural Language Processing Toolkit." *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60, https://doi.org/10.3115/v1/P14-5010. Link Google Scholar
8. Meropol, Neal J., et al. "Barriers to Clinical Trial Participation as Perceived by Oncologists and Patients." *Journal of the National Comprehensive Cancer Network*, vol. 5, no. 8, Sept. 2007, pp. 753-62, https://doi.org/10.6004/jnccn.2007.0067. Accessed 28 July 2021. Link Google Scholar
9. Micheal J. Fox Foundation. *Fox Trial Finder*. *Micheal J. Fox Foundation*, www.michaeljfox.org/trial-finder. Link
10. Ni, Yizhao, et al. "Increasing the Efficiency of Trial-patient Matching: Automated Clinical Trial Eligibility Pre-screening for Pediatric Oncology Patients." *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, 14 Apr. 2015, https://doi.org/10.1186/s12911-015-0149-3. Link Google Scholar
11. Pesaladinne, Nikhil R., et al. *Clinical Trial Selector*. *GirlsComputingLeague*, https://cts.girlscomputingleague.org. Link
12. Pulley, Jill M., et al. "Connecting the Public with Clinical Trial Options: The ResearchMatch Trials Today Tool." *Journal of Clinical and Translational Science*, vol. 2, no. 4, Aug. 2018, pp. 253-57, https://doi.org/10.1017/cts.2018.327. Link Google Scholar

13. Sun, Yingcheng, et al. "The COVID-19 Trial Finder." *Journal of the American Medical Informatics Association*, vol. 28, no. 3, 14 Dec. 2020, pp. 616-21, https://doi.org/10.1093/jamia/ocaa304. Link Google Scholar

14. Thadani, S. R., et al. "Electronic Screening Improves Efficiency in Clinical Trial Recruitment." *Journal of the American Medical Informatics Association*, vol. 16, no. 6, 28 Aug. 2009, pp. 869-73, https://doi.org/10.1197/jamia.M3119. Link Google Scholar

15. Weckstein, Douglas J., et al. "Assessment of Perceived Cost to the Patient and Other Barriers to Clinical Trial Participation." *Journal of Oncology Practice*, vol. 7, no. 5, Sept. 2011, pp. 330-33, https://doi.org/10.1200/JOP.2011.000236. Link Google Scholar