International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-11, Issue-2; Mar-Apr, 2025
Journal Home Page: https://ijaems.com/
DOI: https://dx.doi.org/10.22161/ijaems.112.18

Using Tokenization and Random Forest Models to Predict Pandemic Trial Outcomes

Apeksha Mewani

Received: 25 Mar 2025; Received in revised form: 21 Apr 2025; Accepted: 27 Apr 2025; Available online: 30 Apr 2025

Abstract *– Since the onset of the COVID-19 pandemic, thousands of clinical trials have been launched to evaluate the effectiveness of interventions aimed at preventing or treating the virus. While many of these studies reached completion, a notable proportion were prematurely cessated. Using a comprehensive XML dataset of 5,783 COVID-19 trials registered on ClinicalTrials.gov, we developed a machine learning model to predict whether a trial was likely to be completed or cessated. Our findings, supported by token frequency analysis, highlighted those specific variables, namely the type of intervention and the trial location, played a significant role in distinguishing between outcomes. Trials that included 'hydroxychloroquine' or 'azithromycin' as interventions, and those conducted in locations such as 'France,' were more frequently associated with early cessation, reflecting shifting scientific consensus and regulatory changes over time.*

Keywords *–* ***COVID-19, clinical trials, cessation, interventions, machine learning.***

I. INTRODUCTION

SARS-CoV-2, first identified in Wuhan, China, is the causative agent of the COVID-19 pandemic, which has led to substantial global morbidity and mortality. In response to the emerging crisis, numerous international collaborations accelerated the development of pharmacological interventions to mitigate the health impacts of the pandemic. On January 6, 2020, Chinese authorities alerted the World Health Organization (WHO) about the novel coronavirus, prompting the U.S. Centers for Disease Control and Prevention (CDC) to activate a Level 2 emergency response shortly thereafter [1, 2]. By April 2020, the National Institutes of Health (NIH) launched the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) public-private partnership, designed to prioritize and expedite the evaluation of promising treatments and vaccines through coordinated clinical research [3]. This effort culminated in the release of a comprehensive NIH strategic plan in July 2020 to fast-track the development of effective therapeutics, vaccines, and diagnostic tools [4].

To secure regulatory approval by the U.S. Food and Drug Administration (FDA), these interventions required robust premarketing clinical trial data demonstrating safety and efficacy. As a result, thousands of trials were initiated globally to assess a wide range of COVID-19 prevention and treatment strategies. While many of these trials reached successful completion, others were suspended, terminated, or withdrawn due to evolving evidence, safety concerns, logistical barriers, or shifts in the pandemic landscape. Although several retrospective analyses have explored factors associated with trial completion or cessation, few studies have leveraged predictive modeling to identify trials at high risk of early discontinuation [5]. Given the ongoing need to optimize resource allocation and research planning in the face of future pandemics or emerging health threats, we aim to develop a model that can proactively predict which COVID-19 clinical trials are at greater risk for cessation.

II. MATERIALS AND METHODS

DATA SOURCE

The dataset utilized in this analysis comprises 5,783

COVID-19 clinical trials registered on ClinicalTrials.gov, the most comprehensive registry of private and publicly funded clinical studies conducted globally. Managed by the U.S. National Library of Medicine [6], ClinicalTrials.gov serves as a central resource for accessing detailed information on clinical research and is recognized as the gold standard for trial registries worldwide [7]. The dataset consists of XML-formatted files, with each file representing a single study and capturing extensive trial-specific information.

Key variables extracted from each trial entry include study conditions, sponsoring agency, agency classification, brief and detailed summaries, study status, start date, and participant eligibility criteria. Eligibility information is further delineated into inclusion and exclusion criteria to support refined cohort analysis. Additional metadata includes enrollment size, study phase, and type (e.g., interventional or observational), as well as intervention characteristics such as type and name. Critical elements of study design are also documented, including allocation methods, masking protocols, observation models, time perspectives, primary purposes, endpoint classifications, and geographic locations. This comprehensive structure allows for robust exploration of trial characteristics and enables predictive modeling to assess the factors associated with study cessation.

III. DATA PROCESSING

Data analysis for this project was conducted using Spyder for Python version 4.2.5. The process began with importing all necessary libraries, followed by loading the XML dataset into the working environment. Initial exploration was performed using the .info() command to manually assess the structure of the data. The original dataset contained 5,783 entries and 27 variables. To refine the analysis, we filtered the dataset to include only interventional studies, narrowing the sample to 3,322 trials.

Preprocessing steps included initializing empty lists to store reformatted variables and parameterizing key attributes such as participant age, study phase, trial start and end dates, sponsoring organization, funding source, and reported condition. Regular expressions (regex) were used to clean and standardize text inputs: columns with multiple entries were split using the pipe ("|") delimiter, special characters were removed, and textual numeric values were converted into integers where applicable.

Sponsorship data was recoded into three main categories; government, industry, and other to facilitate subgroup analysis. For age-related data, numerical values were extracted alongside keywords such as "months" and "years" and were classified into three distinct age brackets. The intervention and location fields, which consisted of unstructured free text, underwent additional cleaning procedures. These included removal of special characters, tokenization into individual words, conversion to lowercase, and elimination of stop words—terms that do not add substantive meaning (e.g., "the," "and," "of"). This preprocessing enabled more precise feature extraction for modeling and interpretation.

IV. DATA ANALYSIS

Our investigation was driven by two primary research questions. First, what are the underlying factors that differentiate COVID-19 trial cessation (defined as trials that were terminated, suspended, or withdrawn) from trial completion? Specifically, what contextual or design-related factors may influence whether a trial ceases or completes? Second, can we develop a predictive model to accurately classify COVID-19 trials as either completed or cessated based on available metadata?

To explore the first question, we began by computing descriptive statistics to compare characteristics between completed and cessated trials. We then directed our focus toward two unstructured text variables, location and intervention, which we hypothesized to be contextually significant in determining trial outcomes. We aimed to identify specific words or terms that appeared more frequently in completed trials versus those that were cessated. Using custom Python scripts, we calculated word frequencies and conditional probabilities for each variable and stratified the results by trial status. The distributions of term frequencies were then visualized using the matplotlib library to detect meaningful patterns.

To address our second research question, we constructed a predictive model using the same two variables (location and intervention) as input features. First, we vectorized the cleaned text strings using the Word2Vec algorithm from the gensim library to convert unstructured language into numerical representations. We then implemented a Random Forest classifier from the scikit-learn library. The model was trained on 80% of a balanced dataset, which consisted of 208 randomly selected completed trials and all 208 cessated trials. The remaining 20% of the dataset was used for testing to evaluate the model's performance in predicting whether a trial would complete or cease based on the intervention and location inputs.

V. RESULTS

Table 1: Characteristics of completed and cessated COVID-19 clinical trials.

	Completed	Cessated
Total number of trials included in dataset	460 (13.9%)	191 (5.75%)
Median number of participants	91	0*
Randomized	319 (69.3%)	153 (80.1%)
Non-Randomized	51 (11%)	7 (3.7%)
Open Label	255 (55.4%)	103 (53.9%)

*Most cessated trials were withdrawn, hence the large number of studies with 0 participants.

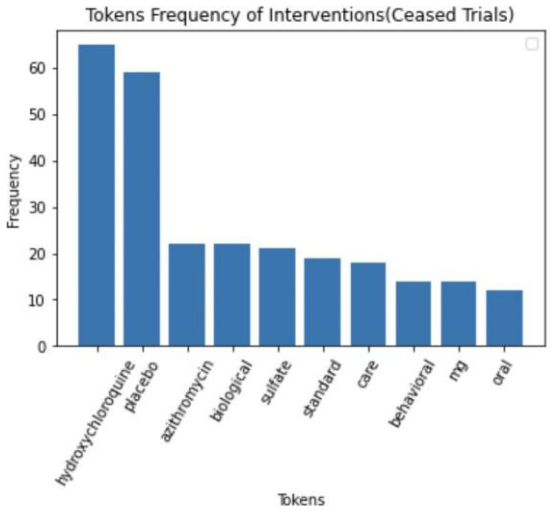
Out of the 5,783 clinical trials in our dataset, 3,322 were classified as interventional in design and were included in our final analysis. Among these, 191 trials (5.75%) were classified as *cessated*, encompassing studies that were withdrawn, terminated, or suspended prior to completion. Specifically, of the 191 cessated trials, 96 (50.3%) were withdrawn, 70 (36.6%) were terminated, and 25 (13.1%) were suspended. A full breakdown comparing completed and cessated trials is provided in Table 1.

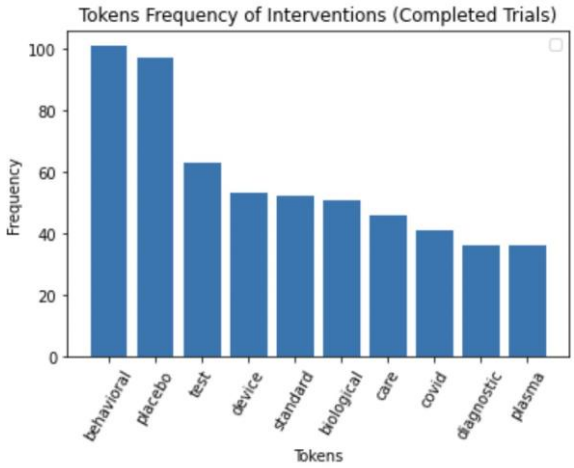
Regarding study design, the majority of cessated trials (144; 75.4%) utilized a parallel assignment model. Other study designs included single group assignment (34 trials; 17.8%), sequential assignment (9 trials; 4.7%), factorial assignment (3 trials; 1.6%), and crossover assignment (1 trial; 0.5%). Enrollment patterns varied significantly between completed and cessated trials. The median number of participants enrolled in cessated trials was 0, with an interquartile range of 0 to 47.5. Notably, 103 of the cessated studies (53.9%) had no participants enrolled at all. When excluding withdrawn studies and examining only suspended and terminated trials, the median enrollment increased to 50 participants, with a range of 17 to 133. In contrast, completed trials had a much higher median enrollment of 91 participants, ranging from 1 to 734,383 participants.

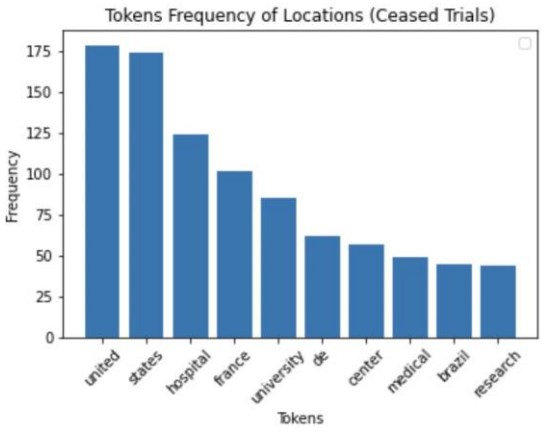
In our frequency and probability analyses of the *intervention* variable, the drug hydroxychloroquine emerged as the most prominent among cessated trials, appearing 65 times across the corpus. This drug was used in approximately 34% of all cessated trials. Similarly, azithromycin was also frequently present in cessated trials, appearing 22 times, corresponding to a 11.5% probability of occurrence in that group. Notably, no specific drugs appeared with comparably high frequency in the *intervention* field of completed trials, suggesting a more heterogeneous distribution of interventions in successful studies.

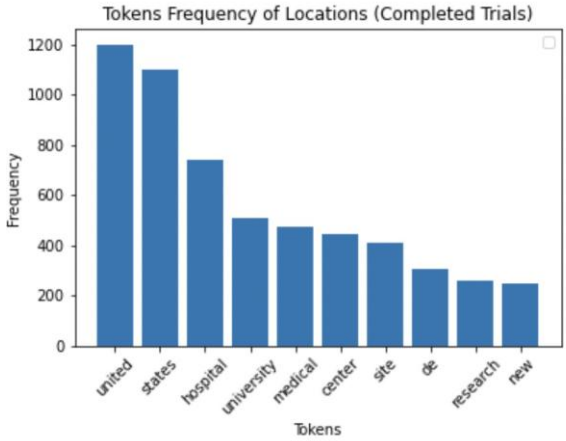
When analyzing the *location* variable, three country names were frequently associated with cessated trials: the United States, France, and Brazil. In contrast, only the United States appeared consistently in completed trials, indicating a possible geographic pattern related to trial discontinuation.

Visualizations of term frequencies and distributions for both the *intervention* and *location* variables, segmented by trial status (completed vs. cessated), are presented below. These figures illustrate the disproportionate presence of certain interventions and locations in discontinued studies, supporting the hypothesis that these unstructured text variables hold predictive value.









From the trained Random Forest classifier, we found the following prediction accuracy measures for the intervention and location variables for the completed and ceased trials: **Location variable**

	Completed	**Ceased**
Precision	0.587	0.857
Recall	0.925	0.409
Fscore	0.718	0.554

Intervention variable

	Completed	**Ceased**
Precision	0.771	0.735
Recall	0.675	0.818
Fscore	0.72	0.774

VI. DISCUSSION

Our analyses both in tokenization frequency and predictive modeling consistently demonstrated that the intervention and location variables were significant in distinguishing between cessated and completed COVID-19 clinical trials. Notably, specific tokens such as hydroxychloroquine and azithromycin within the intervention field, and France within the location field, were disproportionately present in cessated trials. These findings may, in part, reflect the influential role of French microbiologist Dr. Didier Raoult, whose early promotion of hydroxychloroquine (HCQ) and azithromycin (AZM) as a COVID-19 therapy garnered global attention and catalyzed numerous clinical investigations.

Hydroxychloroquine, an antimalarial drug also used in autoimmune conditions like systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA), has mechanisms believed to involve

immunomodulatory and anti-inflammatory effects, though its precise antimicrobial mechanism remains unclear [8]. Early in the pandemic, Dr. Raoult's group published in vitro data suggesting HCQ-AZM synergy was toxic to SARS-CoV-2 [9]. This was followed by a small, non-randomized, open-label clinical trial involving 36 COVID-19 patients, which further fueled enthusiasm for the drug combination [10]. Dr. Raoult's public assertions of success, such as his claim that "we know how to cure the disease"—along with a similar report from China, led to a surge in HCQ-related trials globally. At one point, hydroxychloroquine was the subject of one in every five drug trials registered worldwide [11].

However, the initial promise of HCQ-AZM was not borne out in subsequent, more rigorously designed trials. Accumulating evidence demonstrated not only inconsistent efficacy but also significant risks, particularly the increased likelihood of fatal cardiac arrhythmias when HCQ and AZM were used in combination [8, 12, 13]. Moreover, Raoult's study has been heavily criticized for numerous methodological shortcomings, including a small sample size, poorly defined endpoints, lack of randomization, and absence of blinding [8, 9]. These flaws raised serious concerns about internal validity and potential bias, especially given Dr. Raoult's public advocacy of HCQ. Such methodological limitations may have played a critical role in the early cessation of many similar trials inspired by the initial findings.

Turning to model performance, our predictive analysis further validated the utility of the intervention and location variables. Precision scores defined as the ratio of true positives to all predicted positives, exceeded 50% for both completed and cessated trials using both variables. Recall the ratio of true positives to all actual positives was notably high for completed trials: approximately 90% when using the location variable and 67% when using the intervention variable. For cessated trials, recall was strongest with the intervention variable at 81%, but dropped below 50% when using location as the sole predictor. These results suggest that the intervention variable may be more predictive of cessation than the location variable when assessed using recall.

The F1-score, which balances precision and recall, showed that both variables performed comparably in predicting completed trials. However, for cessated trials, the intervention variable yielded a higher F1-score, reinforcing its greater predictive value in identifying trials at risk for discontinuation.

Future research should expand upon this foundation by exploring additional predictors such as funding source, eligibility criteria, and study phase. Moreover, replicating and refining this predictive model with larger datasets and an expanded range of structured and unstructured features would likely enhance its robustness and utility for real-time risk assessment in trial planning and monitoring.

VII. IMPLICATIONS

The findings of this study have important implications for public health planning and emergency preparedness, particularly in the context of rapidly evolving health crises like the COVID-19 pandemic. By identifying key predictors of clinical trial cessation specifically intervention type and geographic location, this research contributes to a growing body of knowledge that can be leveraged to improve the design, prioritization, and oversight of emergency-related clinical research.

First, the ability to predict which trials are at risk of early termination can help policymakers and funding agencies allocate limited resources more effectively. During pandemics or other public health emergencies, time and funding are often constrained. A predictive framework such as the one developed in this study can inform decision-makers about which proposed studies are more likely to reach completion, thus improving the efficiency of research pipelines and accelerating the delivery of actionable results to clinicians and public health officials.

Second, our findings highlight the potential consequences of early enthusiasm around interventions such as hydroxychloroquine and azithromycin that may lack robust supporting evidence. This underscores the critical need for stronger early-stage vetting of therapeutic candidates and trial protocols, especially when public and political interest can lead to an overconcentration of trials around a limited set of interventions. Strengthening mechanisms for scientific rigor and peer oversight in early-phase trials can reduce redundancy and prevent the rapid proliferation of poorly designed studies during emergencies.

Lastly, incorporating predictive tools into public health infrastructure can support future pandemic preparedness efforts. Such tools can be integrated into clinical trial registries and decision-support systems to flag at-risk studies in real time. This would enable research oversight bodies to intervene early whether by providing additional support, recommending design modifications, or redirecting efforts thereby enhancing the overall resilience and responsiveness of the public health research ecosystem. In sum, predictive modeling of trial outcomes holds promise not only for improving research efficiency but also for ensuring that the scientific response to public health emergencies is timely, evidence-based, and strategically aligned with population health needs.

VIII. LIMITATIONS

This study has several limitations. First, the analysis relied on our own interpretations of variable fields, as the dataset did not include comprehensive metadata or variable descriptions. As a result, some categorizations may have introduced subjective bias. Second, during preprocessing of the intervention and location variables, we excluded symbols and non-English terms to improve consistency and reduce noise; however, this may have resulted in the loss of potentially meaningful information. Third, the predictive model was developed using a limited set of input variables, which increases the risk of overfitting and may limit the model's applicability to other datasets. Finally, the findings are specific to interventional COVID-19 clinical trials and may not be generalizable to observational studies or to trials targeting other diseases or conditions.

IX. CONCLUSION

The results from both our tokenization frequency analysis and predictive modeling were consistent, indicating that the intervention and location variables effectively differentiated between cessated and completed trials. Notably, certain tokens such as hydroxychloroquine and azithromycin under the intervention variable, and France under the location variable were present in cessated trials but absent in completed ones.

REFERENCES

[1] Cham N, Chams S, Badran R et. al. COVID-19: A Multidisciplinary Review. Front Public Health. (2020); 8: 383.

[2] Koopmans M. The novel coronavirus outbreak: what we know and what we don't. Cell (2020); 180 (6): 1034–6.

[3] Collins FS, Stoffels P. Accelerating COVID-19 therapeutic interventions and vaccines (ACTIV): an unprecedented partnership for unprecedented times. JAMA (2020); 323 (24): 2455.

[4] NIH-Wide Strategic Plan for COVID-19 Research 2020.

[5] https://www.nih.gov/sites/default/files/research-training/initiatives/covid-19-strategic-plan/Coronavirus-strategic-plan-20200713.pdf.

[6] He Z, Erdengasileng A, Luo X et. al. How the clinical research community responded to the COVID-19 pandemic: an analysis of the COVID-19 clinical studies in ClinicalTrials.gov. JAMIA Open, 0(0), 2021, 1–12.

[7] ClinicalTrials.gov. History, Policies, and Laws - ClinicalTrials.gov 2020. https://clinicaltrials.gov/ct2/about-site/historyNPRM. Accessed July 10, 2020.

[8] Schwartz LM, Woloshin S, Zheng E, Tse T, Zarin DA. ClinicalTrials. gov

[9] and Drugs@ FDA: a comparison of results reporting for new drug approval trials. Ann Intern Med 2016; 165 (6): 421–30.

[10] Bansai P, Goyal A, Cusick IV A et. al. Hydroxychloroquine: a comprehensive review and its controversial role in coronavirus disease 2019. Annals of Medicine (2019) 53(1): 117-134.

[11] Andreani J, Le Bideau M, Duflot I et. al. In vitro testing of combined hydroxychloroquine and azithromycin on SARSCoV-2 shows synergistic effect. Microbial Pathogenesis (2020) 145:104228.

[12] Guaret P, Lagier J-C, Parola P et. al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. International Journal of Antimicrobial Agents (2020) 56:105949

[13] Sayare S, (2020). He Was a Science Star. Then He Promoted a Questionable Cure for Covid-19.The New York Times
url:
https://www.nytimes.com/2020/05/12/magazine/didier-raoult-hydroxychloroquine.html

[14] Fiolet T, Guihur A, Rebeaud ME et. al. Effect of hydroxychloroquine with or without azithromycin on the mortality of coronavirus disease 2019 (COVID-19) patients: a systematic review and meta-analysis. Clinical Microbiology and Infection (2021)

27:19e27

[15] Tleyjeh IM, Kashour Z, AlDorsay O et. al. Cardiac Toxicity of Chloroquine or Hydroxychloroquine in Patients With COVID-19: A Systematic Review and Meta-regression Analysis. Mayo Clin Proc Inn Qual Out (2021);5(1):137-150.