The YODA Project
Research Proposal Review

The following page contains the final YODA Project review approving this proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Final
(Protocol #: 2022-5013)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	Yes
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Yes
3. Can the proposed research be reasonably addressed using the requested data?	Yes, or it's highly likely
4. Recommendation for this data request:	Approve

Comments:

No additional comments

The YODA Project
Research Proposal Review

Revisions were requested during review of this proposal.
The following pages contain the original YODA Project review and the original submitted proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Revisions Requested
(Protocol #: 2022-5013)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	No
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Yes
3. Can the proposed research be reasonably addressed using the requested data?	Yes, or it's highly likely
4. Recommendation for this data request:	Not Approve

Comments:

1. Please clarify if the ML model be freely available or are there plans to commercialize?
2. Please clarify that the ML model can only identify patients who fail immunosuppressive treatment, not patients who are likely to benefit from HSCT (which remains unproven).
3. Please add more detail to the statistical analysis section in terms of appropriate approaches, including the specific variables for the requested trials.

Principal Investigator

First Name: LUCIANA
Last Name: SOUZA
Degree: master
Primary Affiliation: University of Sao Paulo (USP) - School of Arts, Sciences and Humanities (EACH)
E-mail: lucianassou@usp.br
Phone number: 5511982166744
Address: Rua Arlindo Béttio, 1000 - Ermelino Matarazzo, São Paulo - SP, 03828-000
Rua Arlindo Béttio, 1000 - Ermelino Matarazzo, São Paulo - SP, 03828-000
City: SAO PAULO
State or Province: SP
Zip or Postal Code: 03357-050
Country: Brazil

General Information

Key Personnel (in addition to PI):
First Name: Luciana
Last name: Souza
Degree: Master
Primary Affiliation: University of Sao Paulo - USP
SCOPUS ID:

First Name: Milton Artur
Last name: Ruiz
Degree:
Primary Affiliation: University of Sao Paulo - USP
SCOPUS ID:

Are external grants or funds being used to support this research?: External grants or funds are being used to support this research.
Project Funding Source: CAPES - USP - EACH
How did you learn about the YODA Project?: PubMed

Conflict of Interest

https://yoda.yale.edu/system/files/formcoi.pdf
https://yoda.yale.edu/system/files/formcoi-drmilton.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT00036439 - C0168T37 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
2. NCT00096655 - C0168T46 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis

3. NCT00207675 - C0168T47 - A Randomized, Multicenter, Open-label Study to Evaluate the Safety and Efficacy of Anti-TNF a Chimeric Monoclonal Antibody (Infliximab, REMICADE) in Pediatric Subjects With Moderate to Severe CROHN'S Disease
4. NCT00094458 - C0168T67 - Multicenter, Randomized, Double-Blind, Active Controlled Trial Comparing REMICADE® (infliximab) and REMICADE plus Azathioprine to Azathioprine in the Treatment of Patients with Crohn's Disease Naive to both Immunomodulators and Biologic Therapy (Study of Biologic and Immunomodulator Naive Patients in Crohn's Disease)
5. NCT00487539 - C0524T17 - A Phase 2/3 Multicenter, Randomized, Placebo-controlled, Double blind Study to Evaluate the Safety and Efficacy of Golimumab Induction Therapy, Administered Subcutaneously, in Subjects with Moderately to Severely Active Ulcerative Colitis
6. NCT00207662 - C0168T21 - ACCENT I - A Randomized, Double-blind, Placebo-controlled Trial of Anti-TNFa Chimeric Monoclonal Antibody (Infliximab, Remicade) in the Long-term Treatment of Patients With Moderately to Severely Active Crohn's Disease
7. NCT00207766 - C0168T26 - ACCENT II - A Randomized, Double-blind, Placebo-controlled Trial of Anti-TNF Chimeric Monoclonal Antibody (Infliximab, Remicade) in the Long Term Treatment of Patients With Fistulizing CROHN'S Disease
8. NCT00004941 - C0168T20 - A Placebo-controlled, Repeated-dose Study of Anti-TNF Chimeric Monoclonal Antibody (cA2) in the Treatment of Patients with Enterocutaneous Fistulae as a Complication of Crohn's Disease
9. NCT01190839 - REMICADECRD3001 - Prospective, Multicenter, Randomized, Double-Blind, Placebo-Controlled Trial Comparing REMICADE (Infliximab) and Placebo in the Prevention of Recurrence in Crohn's Disease Patients Undergoing Surgical Resection Who Are at Increased Risk of Recurrence
10. C0168T16 - Efficacy and safety of retreatment with anti-tumor necrosis factor antibody (infliximab) to maintain remission in Crohn's disease.
11. NCT00771667 - C0743T26 - A Phase 2b, Multicenter, Randomized, Double-blind, Placebo-controlled, Parallel Group Study to Evaluate the Efficacy and Safety of Ustekinumab Therapy in Subjects With Moderately to Severely Active Crohn's Disease Previously Treated With TNF Antagonist Therapy
12. NCT01369329 - CNTO1275CRD3001 - A Phase 3, Randomized, Double-blind, Placebo-controlled, Parallel-group, Multicenter Study to Evaluate the Safety and Efficacy of Ustekinumab Induction Therapy in Subjects With Moderately to Severely Active Crohn's Disease Who Have Failed or Are Intolerant to TNF Antagonist Therapy (UNITI-1)
13. NCT01369342 - CNTO1275CRD3002 - A Phase 3, Randomized, Double-blind, Placebo-controlled, Parallel-group, Multicenter Study to Evaluate the Safety and Efficacy of Ustekinumab Induction Therapy in Subjects With Moderately to Severely Active Crohn's Disease (UNITI-2)
14. NCT00488631 - C0524T18 - A Phase 3 Multicenter, Randomized, Placebo-controlled, Double-blind Study to Evaluate the Safety and Efficacy of Golimumab Maintenance Therapy, Administered Subcutaneously, in Subjects With Moderately to Severely Active Ulcerative Colitis
15. NCT01369355 - CNTO1275CRD3003 - A Phase 3, Randomized, Double-blind, Placebo-controlled, Parallel-group, Multicenter Study to Evaluate the Safety and Efficacy of Ustekinumab Maintenance Therapy in Subjects With Moderately to Severely Active Crohn's Disease
16. NCT00488774 - C0524T16 - A Phase 2/3 Multicenter, Randomized, Placebo-controlled, Double-blind Study to Evaluate the Safety and Efficacy of Golimumab Induction Therapy, Administered Intravenously, in Subjects With Moderately to Severely Active Ulcerative Colitis
17. NCT01863771 - CNTO148UCO3001 - A Safety and Effectiveness Study of Golimumab in Japanese Patients With Moderately to Severely Active Ulcerative Colitis
18. NCT01988961 - CNTO148UCO2001 - A Study to Evaluate the Accuracy of a Subset of the Length-109 Probe Set Panel (a Genetic Test) in Predicting Response to Golimumab in Participants With Moderately to Severely Active Ulcerative Colitis
19. NCT02407236 - CNTO1275UCO3001 - A Phase 3, Randomized, Double-blind, Placebo-controlled, Parallel-group, Multicenter Protocol to Evaluate the Safety and Efficacy of Ustekinumab Induction and Maintenance Therapy in Subjects With Moderately to Severely Active Ulcerative Colitis

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

Identification of the Need for Hematopoietic Stem Cell Transplantation in Crohn's Disease Patients: A Ma-chine Learning Approach

Narrative Summary:

Crohn's disease (CD) is a chronic and heterogeneous inflammatory bowel disease that equally affects people of both sexes. Patients with active disease that is refractory to conventional drugs are eligible for alternative treatments. This scenario includes Hematopoietic Stem Cell Transplantation (HSCT). There is currently a Brazilian protocol for treatment with HSCT. This research aims to create a computational resource, using machine learning techniques, capable of identifying the need for hematopoietic stem cell transplantation in patients with Crohn's disease, thus seeking to provide relevant information to the medical team, helping in the strategic decision of assertive treatment.

Scientific Abstract:

Background
Crohn's disease (CD) is a chronic and heterogeneous inflammatory bowel disease that equally affects people of both sexes [1][4]. The etiology is unknown [1][2][4]. The main symptoms include fever, abdominal pain, frequent cramps, diarrhea, generalized fatigue and also weight loss [1][2][4].
The treatment of this disease, as far as it is concerned, aims to obtain its control and stability. Anti-inflammatory, immunosuppressive and immunobiological drugs are commonly used, isolated or associated, depending on each case [4]. Surgical treatment occurs when intestinal complications resulting from obstructions, or fistula perforations, and patients undergoing surgery are at greater risk of undergoing further surgeries [4]. After 10 years of illness, 80% of patients with the disease reported having had some type of disease-related surgery [4]. There is no established treatment to cure the disease [1][2][4]. This scenario includes Hematopoietic Stem Cell Transplantation (HSCT). In the autologous modality, the association of chemoimmunosuppressive drugs modulates the production of bone marrow cells, reprograms the immune system and the reinfused cells act directly on the patients' intestinal lesions, aiming at their healing [4]. Despite this, in general, patients with CD go through a long time of inefficient treatment attempts to meet the criteria of the Brazilian protocol for inclusion in the treatment of HSCT [4][10].
Objective
This research aims to create a computational resource, using machine learning techniques, capable of identifying the need for hematopoietic stem cell transplantation in patients with Crohn's disease, thus seeking to provide relevant information to the medical team, assisting in strategic decision making in the treatment of this disease.
Study Design
This study intends to develop a computational solution based on the testing of different machine learning methods capable of identifying patterns among CD patients. With the creation of such a resource, containing predictors that collaborate in the strategic process of defining unconventional treatment for this disease, we intend to identify patients with a greater probability of being treated with HSCT. With regard to the possible techniques to be tested, and since it is intended to provide subsidies for decision-making by medical teams, it is desirable that the methods used are interpretable, that is, that their decisions can be justified from the point of view of the patient.
Participants
Clinical data from patients who participated in several related studies.
Primary and Secondary Outcome Measure(s)
Our primary outcome measure will be the accuracy of each tested algorithm in predicting the need for HSCT, given the available data, along with a choice for the algorithm that best performs in this task, considering both its performance and explainability. As stated, our measures will focus in commonly used performance metrics for Machine Learning (i.e. f-measure, precision, recall and ROC analysis).
Statistical Analysis
For the evaluation of the implemented techniques, we will use the measures commonly described in the related literature. In this sense, the use of indicators such as Measure F [7], Precision [7] and ROC Curve [8] stands out. In this sense, we will count on the support of the team specialized in DC of the Hospital Beneficência Portuguesa de São José do Rio Preto with collaborator Dr. Milton Arthur Ruiz to evaluate the results.

Brief Project Background and Statement of Project Significance:

This research aims to create a computational resource, using machine learning techniques, capable of identifying the need for hematopoietic stem cell transplantation in patients with Crohn's disease, thus seeking to provide

relevant information to the medical team, assisting in decision making. strategies in the treatment of Crohn's disease.
In this way, reducing the patient's suffering time in an attempt to find an adequate treatment, and consequently minimizing costs with inefficient treatments and hospitalizations. However, the biggest gain that should be highlighted is the recovery of the quality of life of these patients.

Specific Aims of the Project:

This research aims to create a computational resource, using machine learning techniques, capable of identifying the need for hematopoietic stem cell transplantation in patients with Crohn's disease, thus seeking to provide relevant information to the medical team, assisting in strategic decision making in the treatment of this disease.

What is your Study Design?:

Other

What is the purpose of the analysis being proposed? Please select all that apply.

Other

This research intends to use Meta-analysis of data at the participant level to develop a computational model, with the main purpose of finding a pattern among patients with Crohn's Disease.

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

In the current study we will include data from patients who have already been diagnosed with Crohn's Disease. We will not exclude patients, except those who have not yet been diagnosed with Crohn's Disease.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

Our primary outcome measure will be the accuracy of each tested algorithm in predicting the need for HSCT, given the available data, along with a choice for the algorithm that best performs in this task, considering both its performance and explainability. As stated, our measures will focus in commonly used performance metrics for Machine Learning classification (i.e. f-measure, precision, recall and ROC analysis).

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

The Crohn's Disease Activity Index (CDAI) will also be adopted as an attrib-ute used to identify the level of the disease in which the patient is. Depending on this index, the level can be set to Mild, Moderate, Severe, or Remission [3][10]. When the pa-tient has the CDAI index lower than 150, he is in the remission phase of the disease. On the other hand, if the patient has a CDAI greater than 450 he is in the severe phase [3][10].

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

As a next step, the main attributes that will be part of the final machine learning model will be identified. In this sense, we initially considered the use of the following attributes: CRP Level, Platelet Count, Hemoglobin N, Absolute Leukocyte, Monocyte, Lympholiocyte, Eosnophil, Eutrophic N S and B asophilic counts. We will also include some potentially relevant attributes such as Time of Illness, Patient Age, Gender, Current Treatment, T uantity of Treatments and Number of Surgeries. Other attributes may be included, if considered important, depending on the available datasets and the results of statistical correlation tests.

Statistical Analysis Plan:

Clinical and laboratory data from CD patients that we will use will be analyzed with the aid of a computational model best suited for this study.
With the review of the related literature, it was possible to obtain an overview of the computational models whose results would be more easily interpreted, enabling informed decisions with analysis of patient data.
The most recurrent techniques were verified: Decision Tree, Support Vector Machine - SVM, and Bayesian Networks. These will be the techniques that we will initially implement in this research. However, from the continuity of the literature review on the subject, new techniques can be considered.
For the evaluation of the implemented techniques, we will use the measures commonly described in the related literature. In this sense, the use of indicators such as Measure F [7], Precision [7] and ROC Curve [8] stands out, thus reinforcing the idea of ??implementing these techniques in the present study. Other indicators can be used if suggested by the supporters of this study. In this sense, we will count on the support of the team specialized in dc of the Hospital Beneficência Portuguesa de São José do Rio Preto to evaluate the results, having as main collaborator Dr. Milton Arthur Ruiz.
Software Used:
RStudio

Project Timeline:

Activities (Period)
1 Literature review (January 2022 to December 2022)
2 Data acquisition (January 2022 to May 2023)
The. Contact various research sources focused on Crohn's disease (January 2022 through May 2023)
B. Obtain information in a
database for the present research (May 2023)
3 Attribute selection (May 2023)
O. Apply discretization, binarization and normalization (May 2023 and June 2023)
B. Transform and define relevant variables (June 2023)
4 Definition of computational models to be tested,
based on Review of Related Literature (May 2023 and July 2023)
5 Implementation of the developed models (June 2023 and July 2023)
6 Evaluation and analysis of results (July 2023 to October 2023)
7 Publication of Results (November 2023)

Dissemination Plan:

This study contributes to the identification of a more adequate computational model to analyze the data of patients with CD in more depth, offering a resource capable of collaborating with the treatment strategy of this disease.Some specialized magazines on Crohn's Disease showed interest in publishing this study. We will go together with Dr. Milton Artur Ruiz choose the magazine with the greatest impact.

Bibliography:

[1] Protocolo sobre Doença de Crohn. http://conitec.gov.br/images/Protocolos/DoencaCrohn.pdf
[2] Biondo-Simões MLP, Mandelli KK, Pereira MAC, Faturi JL. Opções terapêuticas pa-ra as doenças inflamatórias intestinais: revisão. Rev bras Coloproct, 2003;23(3):172-182
[3] Victoria, Carlos Roberto, Sassak, Ligia Yukie and Nunes, Hélio Rubens de Carva-lho. Incidence and prevalence rates of inflammatory bowel diseases, in midwestern of São Paulo State, Brazil. Arquivos de Gastroenterologia [online]. 2009, v. 46, n. 1 [Ac-cessed 28 June 2021] , pp. 20-25. Available from: <https://doi.org/10.1590/S0004-28032009000100009>. Epub 19 May 2009. ISSN 1678-4219. https://doi.org/10.1590/S0004-28032009000100009.
[4] Milton Artur Ruiz, Roberto Luiz Kaiser Júnior, Luiz Gustavo De Quadros. Terapia ce-lular na doença de Crohn. Editores São José do Rio Preto: DLR Serviços Ltda., 2018. 456 p. 21 x 29.7 cm ISBN: 978-85-53085-00-2
[5] A. Dignass, G. Van Assche, J.O. Lindsay, M. Lémann, J. Söderholm, J.F. Colombel, S. Danese, A. D'Hoore, M. Gassull, F. Gomollón, D.W. Hommes, P. Michetti, C. O'Morain, T. Öresland, A. Windsor, E.F. Stange, S.P.L. Travis, for the European Crohn's and Colitis Organisation (ECCO), The second European evidence-based Consensus on the diag-nosis and management of Crohn's disease: Current management, Journal of Crohn's and Colitis, Volume 4, Issue 1, February 2010, Pages 28–62, https://doi.org/10.1016/j.crohns.2009.12.002
[6] Dias CC, Pereira Rodrigues P, Fernandes S, Portela F, Ministro P, Martins D, et al. (2017) The risk of disabling, surgery and reoperation in Crohn's disease – A decision tree-based approach to prognosis. PLoS ONE 12(2): e0172165. https://doi.org/10.1371/journal.pone.0172165

[7] Waljee AK, Wallace BI, Cohen-Mekelburg S, et al. Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Se-vere Crohn Disease. JAMA Netw Open. 2019;2(5):e193721. doi:10.1001/jamanetworkopen.2019.3721
[8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introdução ao Data Mining - Mine-ração de Dados. Editora Ciência Moderna Ltda.,2009. 900 p. ISBN:978-85-7393-761-9
[9] Veauthier B, Hornecker JR. Crohn's Disease: Diagnosis and Management. Am Fam Physician. 2018 Dec 1;98(11):661-669. PMID: 30485038.
[10] Ruiz MA, Junior RLK, Piron-Ruiz L, Saran PS, Castiglioni L, Quadros LG, Pinho TS, Burt RK. Medical, ethical, and legal aspects of hematopoietic stem cell transplanta-tion for Crohn's disease in Brazil. World J Stem Cells 2020; 12(10): 1113-1123
[11] G. Canino, P. H. Guzzi, G. Tradigo, A. Zhang and P. Veltri, "On the Analysis of Dis-eases and Their Related Geographical Data," in IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 228-237, Jan. 2017, doi: 10.1109/JBHI.2015.2496424.
[12] Chen G and Shen J. Artificial Intelligence Enhances Studies on Inflamatory Bowel Disease. Front. Bioeng. Biotechnol. 9:635764. Jun. 2021, doi: 10.3389/fbioe.2021.635764.
[13] Ayano Kondo, Siyuan Ma, Michelle Y.Y. Lee, Vivian Ortiz, Daniel Traum, Jona-than Schug, Benjamin Wilkins, Natalie A. Terry, Hongzhe Lee, Klaus H. Kaestner. High-ly Multiplexed Image Analysis of Intestinal Tissue Sections in Patients With Inflammato-ry Bowel Disease. Gastroenterology, Volume 161, Issue 6, 2021, Pages 1940-1952, ISSN 0016-5085, https://doi.org/10.1053/j.gastro.2021.08.055.
[14] Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP; International Inflam-matory Bowel Disease Genetics Consortium (IIBDGC). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep. 2019;9(1):10351. Published 2019 Jul 17. doi:10.1038/s41598-019-46649-z.
[15] Fei Yuan, Yu-Hang Zhang, Xiang-Yin Kong, Yu-Dong Cai, "Identification of Can-didate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach", BioMed Research International, vol. 2017, Article ID 5741948, 15 pages, 2017. https://doi.org/10.1155/2017/5741948.
[16] Kowalska-Duplaga, K.; Kapusta, P.; Gosiewski, T.; Sroka-Oleksiak, A.; Ludwig-S?omczy?ska, A.H.; Wo?kow, P.P.; Fyderek, K. Changes in the Intestinal Microbiota Are Seen Following Treatment with Infliximab in Children with Crohn's Disease. J. Clin. Med. 2020, 9, 687. https://doi.org/10.3390/jcm9030687
[17] Seyed Tabib NS, Madgwick M, Sudhakar P. Big data in IBD: big progress for clini-cal practice 2020. Gut 2020;69:1520-1532.
[18] A Chen,Guihua,A Shen, Artificial intelligence and inflammatory bowel disease, Jun 2021, https://www.frontiersin.org/article/10.3389/fbioe.2021.635764.
[19] Alessandro Gonçalves CampolinaI, Rozana Mesquita CiconelliII. Qualidade de vi-da e medidas de utilidade: parâmetros clínicos para as tomadas de decisão em saúde. Universidade Federal de São Paulo (UNIFESP), Departamento de Medicina, Disciplina de Reumatologia. 2006. SciELO - Scientific Electronic Library Online.
[20] André Hideaki Saheki , Rodolfo Sharovsky , Fabio Gagliardi Cozman , Veerle M. H. Coupé. Construção de uma Rede Bayesiana Aplicada ao Diagnóstico de Doenças Car-díacas. 2003. Hospital Universitário - Universidade de São Paulo. http://sites.poli.usp.br/p/fabio.cozman/Publications/Article/saheki-shar...
[21] Fernandes, Fernando Timoteo e Chiavegatto, Alexandre Dias PortoPerspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. Revista Brasileira de Saúde Ocupacional [online]. 2019, v. 44 [Acessado 4 Fe-vereiro 2022] , e13. Disponível em: <https://doi.org/10.1590/2317-6369000019418>. Epub 04 Nov 2019. ISSN 2317-6369. https://doi.org/10.1590/2317-6369000019418.
[22] NLM. Faceli K, Lorena AC, Gama J, Carvalho ACP de LF de. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011
[23] John Paul Mueller, ?Luca Massaron .Aprendizado de Máquina Para Leigos. Editora: books.google.com.br › books – 2019.

Supplementary Material:

https://yoda.yale.edu/sites/default/files/project-yoda.docx