Principal Investigator

First Name: Hanne
Last Name: Vanluchene
Degree: Master of Science in Biomedical Engineering
Primary Affiliation: AZ Delta
E-mail: hanne.vanluchene@azdelta.be
State or Province: West-Vlaanderen
Country: Belgium

General Information

Key Personnel (other than PI):
First Name: Filip
Last name: Baert
Degree: MD gastroenterology
Primary Affiliation: AZ Delta
SCOPUS ID:
Requires Data Access? No

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: Value-Based Partnership Healthcare Collaboration Agreement with Takeda Belgium NV
How did you learn about the YODA Project?: Data Holder (Company)

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/04/SV_57KskaKADT3U9Aq-R_4Kb373buY6ffhDj.pdf
https://yoda.yale.edu/wp-content/uploads/2024/04/SV_57KskaKADT3U9Aq-R_2ilmv0H27rz5lGd.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. NCT00553176 - C0168Z01 - Crohn's Therapy, Resource, Evaluation, and Assessment Tool Registry
2. NCT00036439 - C0168T37 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
3. NCT00096655 - C0168T46 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
4. 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
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. 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
7. 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
8. 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
9. 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)
10. 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
11. 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)

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

Research Proposal

Project Title

Development of an Inflammatory Bowel Disease (IBD) patient population dashboard to monitor and improve care through machine learning

Narrative Summary:

The number of people with Inflammatory Bowel Disease is increasing worldwide. Different therapeutic options, such as biologic treatment, have been developed in the last decades. However treatment often fails. Therefore, this project aims to develop Machine Learning (ML) models as Clinical Decision Support Tool to assist clinicians in predicting biologic treatment outcomes.
To obtain trustable ML models, enough data is needed and therefore data from different clinical trials are requested. The collected data will be merged in one big dataset, based on common datapoints, which will be used for training ML models. As final step, historical data from our local hospital will be used as validation

Scientific Abstract:

Background: The number of people with Inflammatory Bowel Disease (IBD) is increasing worldwide. Biologic medication is an important treatment, but there is often a lack of response. Because of this the load on the healthcare system grows. Therefore, it is important to optimise care delivery, which can be done with the help of clinical decision support tools (CDST).

Objective: This study aims at developing a CDST in the form of machine learning models to support physicians in determining the outcome of biologic treatment.
Study design: Meta-analysis of the parameters of clinical studies that focus on the outcomes of biologic treatment, with the goal to create supervised machine learning models.
Participants: Crohn's disease and ulcerative colitis patients who received biologic treatment and of whom it is registered whether they ended up in response/remission or not. Patients from different clinical studies, as well as patients from our local hospital dataset will be included in the study.
Main outcome measures: Machine learning models that predict whether a patient ended up in response/remission after short-/long-term treatment with a biologic.
Statistical analysis: A combination of data from different clinical studies will be used to create different machine learning models, that will be validated based on parameters like receiver operating characteristic area under the curve (ROC AUC) and F1-score. The first validation happens on a test set that is separated from the training set of the clinical trials datasets. A second validation happens on our local hospital dataset.

Brief Project Background and Statement of Project Significance:

It is estimated that around 0.5% of the population is affected by Inflammatory Bowel Disease (IBD) and the number of patients has been increasing worldwide [1]. IBD, including Ulcerative Colitis (UC) and Crohn’s Disease (CD), is characterised by episodes of diarrhea, abdominal pain and weight loss due to chronic inflammations in the gastrointestinal tract. Different therapeutic options have been developed in the last decades, where biologic treatment plays an important role. However, treatment failure occurs often and there remain a number of refractory patients [2]. The increasing number of IBD patients and the lack of response to biological treatment causes a growing burden on the healthcare system and costs.

Therefore, it is important to optimise care delivery and patient experience. This can be done using a clinical pathway to collect data in a standardized way and visualise it in a clinical dashboard [3]. Adding Clinical Decision Support Tools (CDST) in the dashboard will allow to determine the likelihood a patient will respond to treatment, which helps clinician with treatment decision and will further help to improve care delivery. The main goal of this project is to make Machine Learning (ML) models that can serve as a CDST to predict treatment outcomes of biologics.

Specific Aims of the Project:

The main objective of the study is to develop clinical decision support tools in the form of machine learning models to support physicians in determining the outcome of different treatments with biologics. The second objective is to add these models in an explainable way in a dashboard to facilitate the caregivers in treatment decision.

Study Design:

Meta-analysis (analysis of multiple trials together)

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

Participant-level data meta-analysis

Meta-analysis using data from the YODA Project and other data sources

Develop or refine statistical methods

Research on clinical prediction or risk prediction

Research Methods

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

your study:

Clinical studies that are included for request are searched based on three criteria. The first criterium is that only participants with Crohn’s disease or Ulcerative Colitis are included. The second criterium is that patients must receive a biologic treatment of one of the following medication types: Infliximab, Adalimumab, Golimumab, Vedolizumab or Ustekinumab. The medication types are selected based on the most used biologics in a hospital. A last criterium is that the studies should have response or remission on a specific biologic as outcome. Patients that received placebo treatment during the clinical study will be excluded from this study.

Following clinical trial IDs are included based on the above criteria: NCT00036439, NCT00096655, NCT00553176, NCT00409617, NCT00408629, NCT00385736, NCT00488631, NCT00487539, NCT00488774, NCT02407236, NCT00771667, NCT01369342, NCT01369355, NCT01369329

These clinical trials are requested on the Vivli platform and on the YODA platform. The analysis will be conducted on the secure research environment of Vivli.

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

The outcomes that will be studied from the clinical trials is the outcome of a treatment, whether the patient responded on the treatment or whether the patient ended up in remission because of the treatment. These outcomes will be studied on short-term (6-10 weeks after start treatment) and on long-term (>44 weeks after start treatment).

The study itself will have a machine learning model as primary outcome. To validate the machine learning model, parameters like receiver operating characteristic area under the curve (ROC AUC) and F1-score will be used. No other outcomes than the machine learning model and its performance measures will be considered.

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

The main predictor of this study is whether a treatment will lead to response/remission for a specific patient. Other predictor values could be whether a treatment is still ongoing after 1 year or if complications will occur from the treatment. These other predictor values depend on the available data points from the clinical study.

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

Multiple variables will be studied in the analysis and relevant variables will be included as feature in the machine learning model.

Following datapoints are available in the local dataset from our hospital. When these datapoints are also available in the clinical studies, then they will be added in the analysis for model optimization.

Demographics:
1. Diagnosis (Chron's disease or ulcerative colitis)
2. Age at diagnosis
3. Age at start treatment
4. Gender
5. Family history of IBD
6. Smoking

Disease characteristics:
1. Disease location
2. Extra-intestinal manifestations

3. Disease complications

Labs:
1. Haemoglobin
2. Thrombocytes
3. CRP
4. Albumin

Treatment details:
1. Biosimilars type, medication type
2. Line of biologics treatment
3. Earlier corticosteroid use

Statistical Analysis Plan:

The first step that will be done with the requested data is a statistical description for each dataset separately (mean, standard deviation, minimum, 25% percentile, median, 75% percentile and maximum). Only parameters that occur in all datasets and in the local dataset will be used in the study. Missing data will be handled by imputation. When the number of missing values is too high for a certain parameter, it will be reconsidered if this parameter should be included in the study. The datasets will be brought together in one large dataset without considering independence of datasets, heterogeneity and the potential for confounding as the trial they stem from is irrelevant for this study. A statistical description will be performed on the large dataset as well.

After this, feature selection for the ML model will be performed to only keep the relevant variables. Collinearity and multicollinearity will be studied as they are known to potentially have a negative impact on model stability, explainability and model performance. Collinearity is investigated through a Pearson correlation matrix when data has a normal distribution or through a Spearman correlation matrix when not normal distributed. Normality is examined for the continuous and ordinal variables using the Shapiro-Wilk test. Features with an absolute correlation coefficient higher than 0.6 will be removed. For each remaining feature, multicollinearity will be investigated by computing the Variance Inflation Factor (VIF). Features will be removed to achieve a VIF smaller than 10 for all remaining features. For each statistical test in this study, a level of significance of 0.05 is used.

Now that all features for the ML model are selected, one hot encoding will be performed to make the output of categorical data numerical. After this, the data will be splitted in a train (80%) and test (20%) dataset to be able to validate the results on the test dataset. Before starting to train the ML model, there will be checked whether the dataset is balanced. In case the dataset is unbalanced, upsampling will be performed to increase the number of samples in the minority class to balance out the distribution. After this, different supervised ML models will be created and compared. Following algorithms will be investigated: logistic regression, random forest, support vector classifier, K-nearest neighbours, Gaussian Naïve Bayes, XGBoost and neural networks. To validate the models, following performance measures will be checked on the test dataset: ROC AUC (Receiver Operator Characteristic Area Under the Curve) and F1-score. Once the ML models are trained and tested on the data from the clinical trial studies, another validation will be performed on our local hospital dataset.

Software Used:

Python

Project Timeline:

Following dates are an estimation of the project timeline, it depends on when the data is available and on the project progress.
Anticipated project start date: June 2024
Analysis completion date: March 2025
Date manuscript drafted: April 2025

Date manuscript first submitted: May 2025
Date results reported back to YODA-project: April 2025

Dissemination Plan:

If the ML models reach an acceptable AUC and F1-score, we would like to publish our results in a journal focusing on clinicians of gastroenterology and submit an abstract for a congress (for example the European Crohn's and Colitis Organisation (ECCO) congress).

Journals of interest are:
• JOURNAL of CROHN'S and COLITIS
• CROHN'S & COLITIS 360
• Gastroenterology
• United European Gastroenterology Journal

We would also like to validate the ML models on data from other hospitals, this can be done by sharing the model structure.

Bibliography:

1. Mak JWY, Sun Y, Limsrivilai J, et al. Development of the global inflammatory bowel disease visualization of epidemiology studies in the 21st century (GIVES-21). *BMC Med Res Methodol*. 2023;23(1). doi:10.1186/s12874-023-01944-2
2. Higashiyama M, Hokaria R. New and Emerging Treatments for Inflammatory Bowel Disease. *Digestion*. 2023;104(1):74-81. doi:10.1159/000527422
3. Baert F, Baert D, Pouillon L, Bossuyt P. Quality outcome measures project in IBD: a proof-of-concept benchmarking study in three Belgian IBD units. *Acta Gastroenterol Belg*. 2023;86(4):521-526. doi:10.51821/86.4.11830