Development of a machine-learning model to assess terminal ileum Endoscopic healing in pediatric Crohn's disease from Magnetic Resonance Enterography data

Itai Guez[1], Gili Focht[2], Mary-Louise C. Greer[3], Ruth Cytter-Kuint[4], Li-tal Pratt[5], Denise A. Castro[5], Dan Turner[2], Anne M. Griffiths[6], Moti Freiman[1*]

[1]Technion - Israel Institute of Technology, Haifa, Israel

[2]Shaare Zedek Medical Center, Juliet Keidan Institute of Paediatric Gastroenterology and Nutrition, The Hebrew University of Jerusalem, Jerusalem, Israel

[3] Department of Diagnostic Imaging, The Hospital for Sick Children, Department of Medical Imaging, University of Toronto, Toronto, Canada

[4]Shaare Zedek Medical Center, Paediatric Radiology Unit, Jerusalem, Israel

[5]Kingston Health Sciences Centre, Queen's University, Department of Diagnostic Imaging, Kingston, Canada

[6]The Hospital for Sick Children, Gastroenterology, Hepatology and Nutrition, Toronto, Canada

Corresponding author: Itai Guez, itai.guez@campus.technion.ac.il

Author contributions:

I.G, M.F.: Conception and design, analysis and interpretation of data, drafting the article, and final approval of the version to be published.

G.F, M.L.C.G, D.T: Data collection and interpretation of results, revising the paper critically for important intellectual content, and final approval of the version to be published.

R.C.K, L.P, D.A.C, A.M.G: Data collection and final approval of the version to be published

Conflicts of interest: MCG – AbbVie research grant and honoraria (disclosure, no conflict)

*Taub fellow (supported by the Taub Family Foundation, Technion's program for leaders in Science and Technology)

Funding: None.

Ethics statement: The imageKids study has been approved by the Helsinki committee of the Israeli Ministry of Health. Since all identifying elements related to patient privacy have been removed from the database used in the research presented in this manuscript, it does not need to provide an additional approval of ethics committee.

Data availability statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to restrictions e.g. they contain information that could compromise the privacy of research participants.

Short title: Machine learning model to assess terminal ileum EH

Abbreviations:

EH – Endoscopic healing

CD – Crohn's disease

TI – terminal ileum

MRE – Magnetic Resonance Enterography

SES-CD – Simple Endoscopic Score for Crohn's Disease

MSE – mean-squared-error

AUC – area under the curve

MaRIA - Magnetic Resonance Index of Activity

PICMI - Pediatric Inflammatory Crohn's MRE Index

IBD – inflammatory bowel disease

DWI - diffusion-weighted imaging

GGA – gastroenterologist global assessment

VAS - visual analogue score

RCE – relative contrast enhancement

CRP - C-Reactive Protein

DT – decision tree

FWER – family-wise error rate

ROC – receiver operating characteristic

Abstract:

Background and Aims: Endoscopic healing (EH), is a major treatment goal for Crohn's disease (CD). However, terminal ileum (TI) intubation failure is common, especially in children. We evaluated the added-value of machine-learning models in imputing a TI Simple Endoscopic Score for CD (SES-CD) from Magnetic Resonance Enterography (MRE) data of pediatric CD patients.

Methods: This is a sub-study of the prospective ImageKids study. We developed machine-learning and baseline linear-regression models to predict TI SES-CD score from the Magnetic Resonance Index of Activity (MaRIA) and the Pediatric Inflammatory Crohn's MRE Index (PICMI) variables. We assessed TI SES-CD predictions' accuracy for intubated patients with a stratified 2-fold validation experimental setup, repeated 50 times. We determined clinical impact by imputing TI SES-CD in patients with ileal intubation failure during ileocolonscopy.

Results: A total of 223 children were included (mean age 14.1±2.5 years), of whom 132 had all relevant variables (107 with TI intubation and 25 with TI intubation failure). The combination of a machine-learning model with the PICMI variables achieved the lowest SES-CD prediction error compared to a baseline MaRIA-based linear regression model for the intubated patients (N=107, 11.7 (10.5-12.5) vs. 12.1 (11.4-12.9), $p<0.05$). The PICMI-based models suggested a higher rate of patients with TI disease among the non-intubated patients compared to a baseline MaRIA-based linear regression model (N=25, up to 25/25 (100%) vs. 23/25 (92%)).

Conclusions: Machine-learning models with clinically-relevant variables as input are more accurate than linear-regression models in predicting TI SES-CD and EH when using the same MRE-based variables.

Keywords: Imaging, Endoscopy, Pediatrics.

71

72 ## 1. Introduction

73 Treatment goals for Crohn's disease (CD) have shifted in the past two decades from symptom control towards
74 more objective goals in accordance to a "treat-to-target" philosophy.[1] Specifically, healing of inflammatory
75 lesions[2] defined as endoscopic healing (EH) is currently considered a primary objective treatment goal in
76 clinical trials.[3] The common term "Mucosal healing" has been recently replaced with "endoscopic healing (EH)"
77 (to differentiate from "histological healing").
78 The Simple Endoscopic Score for CD (SES-CD)[4] assessed by ileocolonoscopy is considered the most established
79 endoscopic score for EH evaluation.[5–7] However, ileal intubation failure during ileocolonoscopy occurs in adults
80 in 13% of CD patients[8] and in considerably higher rates of up to 20-25% of pediatric CD patients.[3,9] This is
81 especially concerning as CD is mostly prevalent in the small bowel with more than 50% of CD patients having
82 terminal ileum (TI) involvement.[10] Consequently, a patient with ileal intubation failure during ileocolonoscopy
83 can be falsely classified as achieving EH, posing a significant challenge for clinical trials.

84 Magnetic resonance enterography (MRE) has emerged as an imaging modality for evaluating transmural
85 healing.[2,11] The most common MRE index in adults of CD activity is the Magnetic Resonance Index of Activity
86 (MaRIA).[12] Recently, the Pediatric Inflammatory Crohn's MRE Index (PICMI) has been developed and validated
87 specifically to assess inflammatory activity in pediatric CD patients. It does not require rectal enema or
88 gadolinium-based contrast agents, and it uses the sum of all affected segments in the entire small and large
89 bowel, thereby reflecting the more extensive nature of disease in CD patients.[13]

90 Imputation of the TI SES-CD subscore from MRE data has the potential to improve overall EH assessment in
91 non-intubated patients. Weiss et al previously developed a linear regression model for imputing the TI SES-CD
92 subscore from MRE features with data collected as part of the ImageKids study.[3,9] However, the model had
93 only a moderate correlation with the TI subscore.[9,14] In recent years, non-linear machine-learning models

demonstrated their potential in improving inflammatory bowel disease (IBD) assessment in various clinical applications by leveraging complex non-linear correlations between the input variables and outcomes.[12,15,16]

The goals of this study were to evaluate the added-value of non-linear machine-learning models in imputing TI SES-CD from MRE data of pediatric CD patients compared to linear-regression models and to determine the most predictive set of MRE-based variables for EH prediction. The machine-learning-based models will have the potential to aid physicians in avoiding misdiagnoses, improving assessment in clinical trials, and ultimately guiding appropriate treatment options for CD patients. The presented models to impute TI SES-CD from MRE variables are available to the community through a dedicated website at: https://tcml-bme.github.io/ML_SESCD.html

104 ## 2. Methods

105 ### 2.1 Study Design

106 This is a sub-study of the multicenter ImageKids study (NCT01881490[17]) in which a total of 240 children who
107 fulfilled the eligibility criteria of age 6 to 18 years and an established diagnosis of CD with ileocolonoscopy
108 followed by an MRE within 14 days without any change in treatment were enrolled. MRE sequences and their
109 scoring system were standardized across centers.[18] Each protocol included a localizer sequence, a motility
110 sequence in the coronal plane, followed by a series of coronal and/or axial rapid T2-weighted, T1-weighted
111 gradient echo sequences and diffusion-weighted imaging (DWI). T1-weighted sequences were performed pre-
112 and post-intravenous gadolinium-based contrast agent (GBCA) injection. An intravenous antispasmodic agent
113 (glucagon or hyoscine butylbromide) was administered following the motility sequence. MRE was performed
114 without enema as it is less feasible in pediatrics. Data were collected from 22 pediatric IBD centers in North
115 America, Europe, Australia and Israel.[18]

116 ### 2.2 Endoscopic report and gastroenterologist assessment

117 The SES-CD evaluates four items (ulcer size, proportion of the surface area that is ulcerated, proportion of the
118 surface area affected, narrowing) by scoring each item on an ordinal scale from 0 to 3 in five bowel segments
119 (TI, right colon, transverse colon, left colon, rectum), where higher score indicates a more severe disease
120 defined by higher percentage of mucosal inflammation. The maximum score in a passable segment is 11 while
121 a score of 12 can be obtained only in the last evaluated segment when there is impassable stenosis. The total
122 SES-CD score, which is the sum of the five segments scores, ranges from 0 to 56. No adjustment is made to
123 account for non-visualized segments (due to resection, technical difficulty or stenosis).[4] The SES-CD items were
124 scored by a gastroenterologist during ileocolonoscopy for each bowel segment. Gastroenterologist global
125 assessment (GGA) of the degree of inflammation was scored on a 100-mm visual analogue score (VAS) and on
126 a Likert scale from 1 (i.e. Complete deep remission) to 5 (i.e. Severe disease activity).

127 **2.3 MRE and radiologist assessment**

128 Each MRE was read three times by independent radiologists highly experienced in pediatric IBD (an on-site
129 radiologist and two central radiologists using an explicit case report form). The radiologists were blinded to the
130 clinical and endoscopic data. Radiologists completed, among others, the Radiologist global assessment (RGA)
131 as well as the MaRIA variables,[19] the PICMI variables,[13] and the length of the diseased segments. The TI was
132 defined in the Imagekids study as the first 10cm segment of the ileum measured from the ileocecal valve. The
133 PICMI variables, the MaRIA variables and the RGA score were read by two independent radiologists and in case
134 of large discrepancy between readings, a third reading was added. Table supp1 in the supplementary
135 materials summarizes the variables for each index.

136 **2.4. Development of ML algorithms**

137 We developed two non-linear machine-learning-based models using the random-forest algorithm[20] to predict
138 EH from MRE each using a different set of variables. The first model used the MaRIA variables, and the second
139 used the PICMI variables to predict EH. In addition, we constructed two multiple-linear-regression models
140 each using the same two sets of variables as input features, and a linear-regression model based solely on the
141 original MaRIA index to serve as baseline methods.

142 For development and validation purposes we used only patients with an available TI SES-CD score which had all
143 MRE variables of both indices. The segment length was not used as a variable for the development as it is not
144 part of MaRIA and PICMI indices.
145 Table supp4 summarizes the full parameter list used for configuration and optimization of the machine
146 learning algorithms.

147 **2.5 Statistics**

148 We used a total of 100 folds (stratified 2-fold validation, repeated 50 times) in order to train each algorithm
149 version with a random set of patients used for derivation and validation. We assessed the added-value of the
150 machine-learning-based models in imputing TI SES-CD values by comparing the accuracy of the imputed TI-SES-
151 CD predicted machine-learning-based models based on the MaRIA and PICMI variables to the predictions
152 made by their counterpart multiple-linear regression models and to the linear regression model with the
153 MaRIA index. We used the mean-squared-error (MSE) from the reference endoscopic TI SES-CD as the
154 measure of accuracy. We also determined which set of variables (PICMI, or MaRIA) provides the most accurate
155 predictions in comparison to the reference endoscopic TI SES-CD measurements.

156 We used the Wilcoxon non-parametric test[21] with Bonferroni correction to control the family-wise error rate
157 (FWER)[22] to determine whether the median of validation MSE differed between two given models.
158 In addition, we used receiver operating characteristic (ROC) curves to evaluate the capacity of the different
159 models' predictions of TI SES-CD to distinguish between patients with and without EH in comparison with
160 reference EH assessments based on endoscopic TI SES-CD. We defined reference EH as SES-CD<3.[23] The
161 predicted SES-CD was normalized to a 0-1 range by dividing it by the maximum possible SES-CD value of 12.
162 We determined the whether the difference in area under the ROC curve (AUC) of the different models is
163 statistically significant with the Delong's test.[24] The models were written in Python [version 3.6.4] using the
164 open-source Scipy [version 1.5.4] library and the open-source Statsmodels [version 0.11.0] library. Models ran
165 on an Intel i3 CPU.

166 Finally, we assessed the clinical added-value of the models in imputing SES-CD for non-intubated patients. We
167 trained the different models on all available data from patients with endoscopic TI SES-CD and imputed SES-CD
168 from the MRE data for the non-intubated patients.

169

170 ## 3. Results

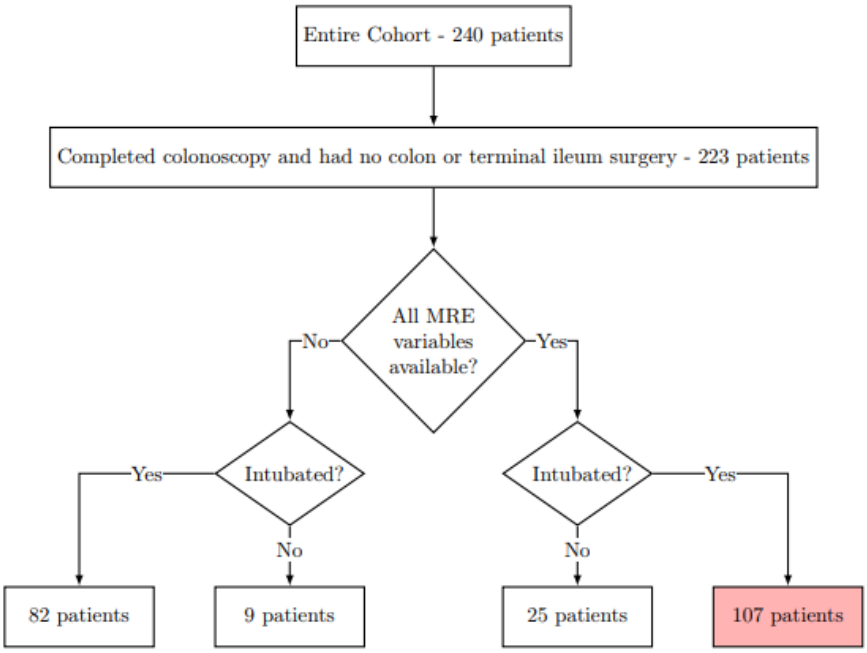
171 ### 3.1 Study population

172 Of the 240 children included in the ImageKids study, 17 were excluded (5 due to incomplete colonoscopy and
173 12 due to prior bowel surgery). The ileal intubation failure rate among the 223 included children was 15%
174 (n=34) and 18% (n=43) among the total cohort. There was no statistically significant difference (Mann-Whitney
175 non-parametric test) in all variables between the intubated and non-intubated groups except for C-Reactive
176 Protein (CRP). GGA-VAS and the ratio of patients with disease limited to TI was higher in the non-intubated but
177 not statistically significant. (Table supp2 in supplementary materials).

178 ### 3.2 Patient-level indices: MRE report

179 A total of 132 children who had all relevant PICMI and MaRIA variables were included in this sub-study. 107 of
180 them had a TI SES-CD score (intubated group), while 25 did not have TI SES-CD scores due to ileal intubation
181 failure during ileocolonoscopy (non-intubated group). Figure 1 summarizes the patients selection process for
182 the study. Table supp3 summarizes the demographic and variables values for the patients included in this sub-
183 study.
184 There were no differences in the population basic characteristics between the intubated and non-intubated
185 groups except for the PICMI score (21.5±10.43 vs. 26.52±7.63, p=0.015). Although not statistically significant,
186 GGA-VAS and CRP had a higher mean score in the group with no ileal intubation (GGA-VAS: 37 (20-60) vs. 40
187 (22-70), p=0.22, and CRP: 7.3 (1.8-25.8) vs. 12.65 (4.98-24.68), p=0.14).

188



189 Figure 1: Patients selection from the entire cohort. MRE: Magnetic Resonance Enterography

190 **3.3 SES-CD prediction from MRE on intubated patients**

191 The machine-learning-based models reduced the MSE of the SES-CD predictions compared to their multiple-
192 linear-regression counterparts. The PICMI-based machine-learning model (RF-PICMI) achieved MSE of 11.8±1.5
193 while its multiple-linear-regression counterpart yielded a higher MSE of 12.1±1.5. Similarly, the MaRIA-based
194 machine-learning model (RF-MaRIA) achieved an MSE of 12.3±1.6 while its multiple-linear-regression
195 counterpart yielded a higher MSE of 12.5±1.6 (table 1).

196 The PICMI-based models reduced the MSE over MARIA-based models for both the machine-learning-based
197 models (RF-PICMI 11.7 (10.5-12.5) vs. RF-MaRIA 12.1 (11.1-13.3), p<1e-5) and the multiple-linear-regression
198 models (MLR-PICMI 11.7 (11.1-13) vs. MLR-MaRIA 12.4 (11.6-13), p<0.05). Finally, the PICMI-based machine-
199 learning model (RF-PICMI) achieved the lowest MSE compared to the baseline linear-regression model with the
200 MaRIA score (RF-PICMI 11.7 (10.5-12.5) vs. LR-MaRIA 12.1 (11.4-12.9), p<0.05) (table 1).

MSE distribution	LR-MaRIA	RF-MaRIA	RF-PICMI	MLR-MaRIA	MLR-PICMI
Mean ± SD	12.2±1.1	12.3±1.6	11.8±1.5	12.5±1.6	12.1±1.5
Median	12.1	12.1	11.7	12.4	11.7
Q1	11.4	11.1	10.5	11.6	11.1
Q3	12.9	13.3	12.5	13	13

Table 1: Validation Mean Squared Error (MSE) distribution statistics over the folds per model. MSE: Mean-squared-error, LR: Linear regression, RF: random-forest, MLR: multiple-linear-regression.

Fig. 2 depicts the MSE distribution of the validation group over the different folds.

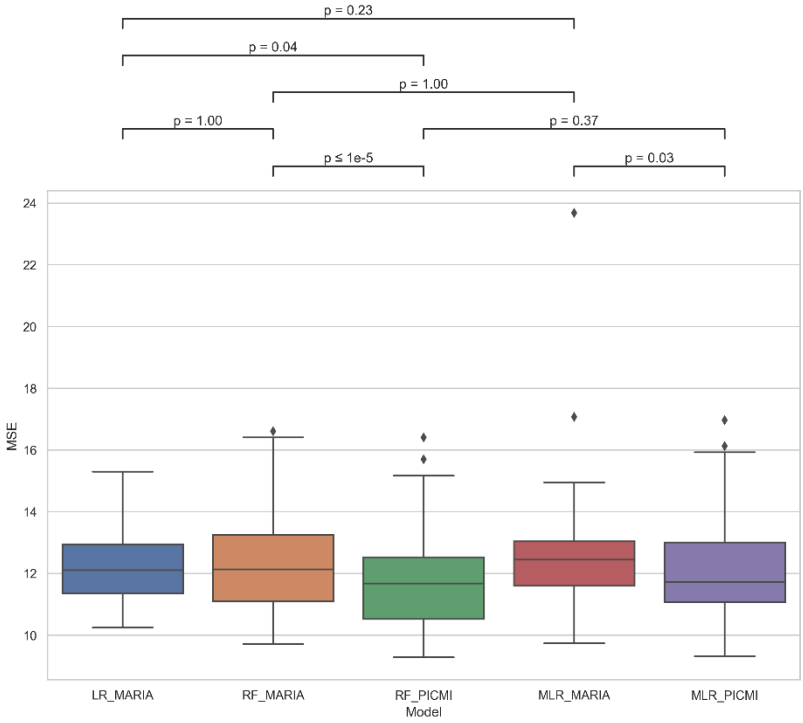


Figure 2: Mean-squared-error (MSE) distribution over the validation folds

Figure 3 presents the averaged ROC curves of the different models' SES-CD predictions in differentiating between patients with and without EH according to their reference ileocolonscopy-based SES-CD score. The

machine-learning-based model with the PICMI variables (RF-PICMI) was able to achieve a similar performance compared to the linear-regression model with the MaRIA index (0.77±0.05 vs. 0.77±0.04, p=0.22, DeLong's test). The MaRIA-based linear-regression model is dependent upon a post-gadolinium T1-weighted sequence.

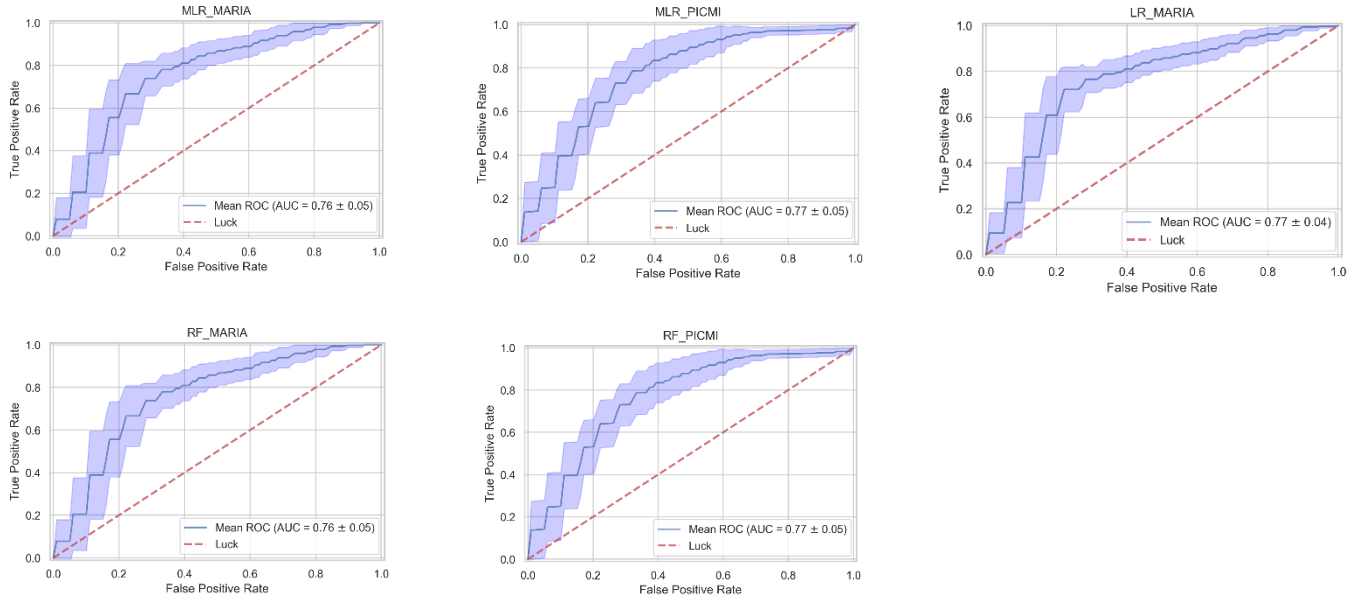


Figure 3: Average ROC curves over the different folds for all models

3.4 SES-CD imputation from MRE on the non-intubated patients

We derived final models on the entire development set of 107 patients and imputed the SES-CD scores for the 25 patients with ileal intubation failure who had all MRE variables. The PICMI-based models predicted higher percentage of patients with ileal disease compared to the baseline linear regression model with the MaRIA index (up to 25 out of 25 (100%) vs. 23 of 25 (92%)). Without imputing TI SES-CD, such non-intubated patients would have been falsely classified as patients without terminal-ileum disease based on colonoscopy only.

MSE distribution	LR-MaRIA	RF-MaRIA	RF-PICMI	MLR-MaRIA	MLR-PICMI
Mean ± SD	4.48±1.5	5±1.9	5±1.3	4.56±1.53	5.36±1.33
Median	5	5	5	5	5
Q1	3	5	5	4	5
Q3	5	6	6	6	6
SES-CD ≥3 (%)	23(92)	21(84)	24(96)	23(92)	25(100)

Table 2: Imputed SES-CD scores for the non-intubated patients. MSE: mean-squared-error, LR: Linear-regression, RF: Random-forest, MLR: Multiple-linear-regression.

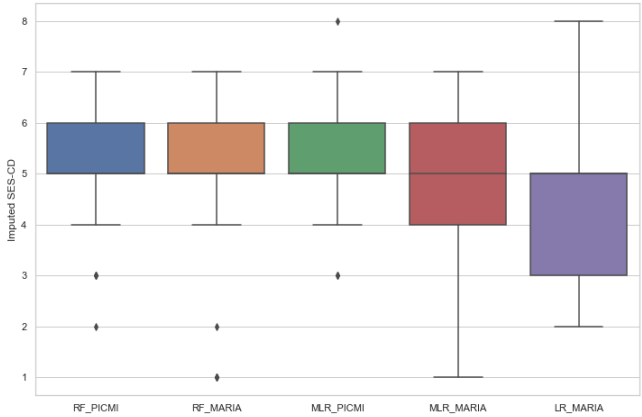


Figure 4: Imputed SES-CD distribution on the non-intubated patients

233 **4. Discussion**

234 Our results demonstrate the potential of machine-learning-based methods in improving MRE-based indices for
235 EH assessment in pediatric CD patients. The combination of the most informative MRE variables (PICMI) and
236 the machine-learning-based model achieved the lowest SES-CD prediction error compared to baseline linear
237 and multiple linear regression models. Further, imputing SES-CD from MRE data for patients with TI intubation
238 failure during ileocolonoscopy using these models suggested a higher rate of patients with TI disease who
239 otherwise were classified as patients without TI disease. Such information might impact interpretation of
240 results in clinical trials.

241 Recently, Weiss et al imputed TI SES-CD from MRE data on pediatric CD patients with a simple linear regression
242 model with the MRE-based MaRIA index as input variable with moderate correlation to the ileo-colonoscopy-
243 based SES-CD score.[14,23]

244 Recently, non-linear machine-learning models such as the Random-Forest model[20] demonstrated their ability
245 to capture complex, non-linear correlations between the input variables and the expected outcomes in
246 multiple domains including in IBD.[12,15,16,26–28] Specifically, Dhaliwal et al. used a machine-learning model to
247 distinguish pediatric colonic IBD subtypes based on clinical, endoscopic, radiologic, and histologic data[29] and
248 Waljee et al. demonstrated the added value of a machine-learning model to predict disease course from
249 demographic data and clinical records compared to a linear-regression model.[30]

250 In this study we hypothesized that the moderate correlation to the TI SES-CD achieved by MRE indices may be
251 attributed, at least in part, to the utilization of classical linear models to characterize the correlation between
252 the MRE-based variables and EH which limited their ability to determine the potentially complex, non-linear
253 correlation. Our results suggest that non- linear machine-learning models can improve the characterization of
254 the complex non-linear correlation between the MRE-based variables and TI SES-CD compared to linear
255 models.

The better performance of the machine-learning-based models compared to their multiple-linear-regression counterpart was expected as the machine-learning-based models have a better capability to characterize complex, non-linear, correlations between input variables and outcomes compared to linear models. In addition, machine-learning-based models are known to have an improved ability to successfully account for noisy data entries that commonly present in medical applications.[28] It is important to note that machine-learning-based models lack the ability to extrapolate predictions beyond the data range used for their development compared to their linear counterparts.[31] However, the lack of extrapolation ability is not a limitation specifically for imputing TI SES-CD since the dataset had rich enough values to cover the entire breadth of scores.

The better performance of the linear-regression model with the MaRIA index compared to its more complex counterpart, the multiple-linear-regression with the MaRIA variables might be explained by the fact that the original MaRIA index coefficients were derived in a more robust way compared to our stratified k-fold experiments.

Finally, the importance of developing pediatric specific MRE-indices was demonstrated by the improved performance of the PICMI-based models compared to the MaRIA-based models. Though previous research questioned the need to develop a specific MRE index of activity for the pediatric population,[2] our results indicate its importance.

The machine-learning-based model with PICMI variables used a DWI sequence as a replacement for the post-gadolinium T1-weighted sequence, an advantage given some evidence this sequence can be omitted on MRE in children to reduce use of GBCA, although not yet standard practice.[25] The better performance of the PICMI-based models compared to that of the MaRIA-based models suggests that the diagnostic accuracy of DWI, included in PICMI and not MaRIA is not overly estimated as was previously hypothesized.[32]

There are several limitations to our study. First, the machine-learning models in this study were developed

with PICMI and MaRIA variables obtained by expert central reading. Further, the PICMI and MaRIA variables were obtained from MRE data acquired with an optimized and standardized acquisition protocol. Models' predictions using PICMI and MARIA variables obtained by radiologists' readings of MRE data that was not necessarily standardized according to the ImageKids MRE protocol should be interpreted with caution. Second, the development of machine-learning models are assumed to require more data compared to the development of linear models due to their complexity. Therefore, the presented models' performance might underestimate the added-value of the machine-learning models. Third, CD is a transmural disease rather than a mucosal disease. It is not clear yet whether mucosal or transmural healing should be the primary treatment goal.[33] While MRE has the ability to detect CD lesions beyond the mucosa, its ability to assess low severity mucosal lesions is much more limited.[2] Therefore, combined assessment with MRE and ileocolonscopy is ideally required to fully assess the extent of the disease.

Last, the models developed in this research are using a set of MRE variables intended to assess transmural healing and not EH. Mesenteric variables for example are indicative of transmural healing but not EH as they measure inflammation in tissues far from the mucosa. The length of an inflamed segment which is proven to be an important indicator of the total burden of the disease in a certain bowel segment is included in magnetic resonance enterography global score (MEGS) variables but not in the MaRIA and PICMI variables.[34] A future study is required to optimally select MRE variables for the sole purpose of mucosa assessment. Tailored mucosal variable selection can further improve SES-CD prediction accuracy.

In summary, our study demonstrates the important role of the combination of non-linear machine-learning models and the selection of specific MRE-based variables to improve our ability to impute TI SES-CD score from MRE data. Such machine-learning methods have the potential to improve MRE-based IBD assessment in additional applications.

[Type here]

5. Acknowledgements

The ImageKids study was supported by an educational grant from AbbVie who were not involved in any part of the study design, conduct, analysis or writing.

[Type here]

321 **6. References**

322 1. L. Peyrin-Biroulet W. Sandborn BES et al. Selecting Therapeutic Targets in Inflammatory Bowel Disease
323 (STRIDE): determining therapeutic goals for treat-to-target. *Am J Gastroenterol*. 2015;110(1324-1338).

324 2. J. Rimola JP. Is the Objective of Treatment for Crohn's Disease Mucosal or Transmural Healing? *Clin*
325 *Gastroenterol Hepatol*. 2018;16(7):1037-1039.

326 3. Turner D, Griffiths AM, Wilson D, et al. Designing clinical trials in paediatric inflammatory bowel
327 diseases: a PIBDnet commentary. *Gut*. 2020;69(1):32-41.

328 4. Daperno M D'Haens G VAG et al. Development and validation of a new, simplified endoscopic activity
329 score for Crohn's disease: the SES-CD. *Gastrointest Endosc*. 2004;60(505–512).

330 5. et al. DZS. Effect of Standardised Scoring Conventions on Inter-rater Reliability in the Endoscopic
331 Evaluation of Crohn's Disease. *J Crohn's Colitis*. 2016;1006(10-14).

332 6. Khanna R Bouguen G FBG et al. A systematic review of measurement of endoscopic disease activity and
333 mucosal healing in Crohn's disease: recommendations for clinical trial design. *Inflamm Bowel Dis*.
334 2014;20:1850-1861.

335 7. Khanna R Zou G DG et al. Reliability among central readers in the evaluation of endoscopic findings
336 from patients with Crohn's disease. *Gut*. 65(1119-1125, year={2016},).

337 8. SH J, KJ L, YB K, HC K, SJ S, JY C. Diagnostic value of terminal ileum intubation during colonoscopy. *J*
338 *Gastroenterol Hepatol*. 2008;23(1):51-55. doi:10.1111/J.1440-1746.2007.05151.X

339 9. Weiss B, Turner D, Griffiths A, et al. Simple Endoscopic Score of Crohn Disease and Magnetic Resonance
340 Enterography in Children: Report from ImageKids Study. *J Pediatr Gastroenterol Nutr*. 2019;69(4):461-
341 465. doi:10.1097/MPG.0000000000002404

342 10. Gajendran M Loganathan P CAPHJG. A comprehensive review and update on Crohn's disease. *Dis Mon*.
343 2018;64(2)(20-57).

344 11. Walsh AJ BR V, SP. T. Current best practice for disease activity assessment in IBD. *Nat Rev Gastroenterol*
345 *Hepatol*. 2016;13(1):567–579.

346 12. Olivera P, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: a look into the future. *Nat Rev*
347 *Gastroenterol \& Hepatol*. 2019;16(5):312-321.

348 13. Turner D Gavish M FG et al. Development of the Pediatric Inflammatory Crohn's MRE Index (PICMI)-
349 results from the ImageKids Study. *J Crohn's Colitis*. 2018;12(1):158-159.

350 14. Rozendorn N. AMMERAKUKE. A review of magnetic resonance enterography-based indices for
351 quantification of Crohn's disease inflammation. *Ther Adv Gastroenterol*. 2018;11(1):1-21.

352 15. Le Berre C, Sandborn WJ, Aridhi S, et al. Application of artificial intelligence to gastroenterology and
353 hepatology. *Gastroenterology*. 2020;158(1):76-94.

354 16. Seyed Tabib NS, Madgwick M, Sudhakar P, Verstockt B, Korcsmaros T, Vermeire S. Big data in IBD: big
355 progress for clinical practice. *Gut*. 2020;69(8):1520-1532.

356 17. https://clinicaltrials.gov/ct2/show/study/NCT01881490.

357 18. et al. W-N, study group I. Associations Among Mucosal and Transmural Healing and Fecal Level of
358 Calprotectin in Children With Crohn's Disease. *Clin Gastroenterol Hepatol*. 2018;16(7)(1089-1097).

359 19. Rimola J Ordás I RS et al. Magnetic resonance imaging for evaluation of Crohn's disease. *Inflamm Bowel*
360 *Dis*. 2011;17(759–68).

361 20. Breiman L. Random Forests. *Mach Learn*. 2001;45(5–32).

362 21. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bull*. 1945;1(6):80-83.

[Type here]

363 22. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*.
364 1988;75(4):800-802.

365 23. Weiss B Turner D GAWTH-SI et al. Simple endoscopic score of Crohn disease and magnetic resonance
366 enterography in children: report from ImageKids Study. *J Pediatr Gastroenterol Nutr*. 2019;69(461-465).

367 24. DeLong ER DeLong DM C-PDL. Comparing the areas under two or more correlated receiver operating
368 characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

369 25. Moy MP, Sauk J, Gee MS. The role of MR enterography in assessing Crohn's disease activity and
370 treatment response. *Gastroenterol Res Pract*. 2016;2016. doi:10.1155/2016/8168695

371 26. Klang E, Grinman A, Soffer S, et al. Automated Detection of Crohn's Disease Intestinal Strictures on
372 Capsule Endoscopy Images Using Deep Neural Networks. *J Crohn's Colitis*. 2020;15(5):749-756.

373 27. Klang E Barash Y MRY et al. Deep learning algorithms for automated detection of Crohn's disease ulcers
374 by video capsule endoscopy. *Gastrointest Endosc*. 2020;91(606-613).

375 28. Y. Q. Random Forest for Bioinformatics. *Ensemble Mach Learn*. Published online 2012.

376 29. Dhaliwal J, Erdman L, Drysdale E, et al. Accurate Classification of Pediatric Colonic Inflammatory Bowel
377 Disease Subtype Using a Random Forest Machine Learning Classifier. *J Pediatr Gastroenterol Nutr*.
378 2021;72(2):262-269.

379 30. Waljee AK, Lipson R, Wiitala WL, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in
380 Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm Bowel Dis*. 2017;24(1):45-53.

381 31. Hengl T, Nussbaum M, Wright M, Heuvelink G, Graeler B. Random forest as a generic framework for
382 predictive modeling of spatial and spatio-temporal variables. *PeerJ*. 2018;6:e5518.

383 32. et. el JR. Development and Validation of a Simplified Magnetic Resonance Index of Activity for Crohn's

384 Disease. *Gastroenterology*. 2019;157(2):432-439.e1.

385 33. ED. S. Treat-to-target in Crohn's disease: Will transmural healing become a therapeutic endpoint?
386 *World J Clin Cases*. 2018;6(12):501-513.

387 34. Zheng X, Li M, Wu Y, et al. Assessment of pediatric Crohn's disease activity: validation of the magnetic
388 resonance enterography global score (MEGS) against endoscopic activity score (SES-CD). *Abdom Radiol*.
389 2020;45(11):3653-3661. doi:10.1007/s00261-020-02590-8

390

391

392

393

394

395

396

397 **7. Supplementary material**

Variable	PICMI	MaRIA	Range	Definition
Wall-thickness	V	V	Positive Number (mm)	The maximum wall thickness in each anatomical division across all sequences
Ulcers	V	V	No - 0, Yes - 1	Deep depressions in the mucosal surface of a thickened segment
Mesenteric edema	V	-	No - 0, Yes - 1	Increased signal intensity of the mesentery on T2-weighted sequences judged subjectively relative to unaffected mesentery
Mural edema	-	V	No - 0, Yes - 1	T2 hyperintensity relative to psoas muscle
Relative Contrast Enhancement (RCE)	-	V	Unbounded number	Increased signal intensity of the bowel wall (WSI) on post-contrast T1-weighted sequences, judged subjectively relative to adjacent normal bowel loops and compared with pre-contrast T1-weighted images. WSI is calculated by the average of three wall enhancement measurements. RCE = [(WSI post-gadolinium – WSI pre-gadolinium)/(WSI pre-gadolinium)]* 100* (SD noise pre-gadolinium / SD noise post-gadolinium), where SD noise pre-gadolinium corresponds to the average of three SD of the signal intensity measured outside of the body before gadolinium injection, and SD noise post-gadolinium corresponds to the SD of the same noise after gadolinium administration.
DWI	V	-	0 - None/Mild, 1 – Moderate/Marked	Increased signal on b images (trace images) with signal drop-out on the apparent diffusion coefficient (ADC) map
Comb sign	V	-	No - 0, Yes - 1	Engorgement of the vasa recta
Length of site	-	-	Positive Number (0-10cm)	Radiologist's linear measurement of length of bowel affected

398

399 Table supp1: Variables of the MRE-based indices

400

401

	Entire group n=223	TI Intubation n=189	TI not intubated n=34	p
Male (%)	128 (57)	107 (56)	21 (62)	0.28
Disease duration, years	2.2 (0.3-4.4)	2.2 (0.3-4.1)	2.8 (0-5.47)	0.22
Age at enrollment, years	14.1±2.5	14.2±2.5	14±2.8	0.33
L1 (TI and cecum) *	57 (25)	45 (24)	12 (35)	0.07
L2 (Colonic)*	33(15)	30(16)	3(9)	0.14
L3 (Ileocolonic)*	128 (57)	110 (57)	18 (53)	0.28
L4 (Isolated upper GI)*	4(2)	4(2)	0(0)	0.19
Fecal Calprotectin (FC)	621.7 (179-1268) available in 184	630 (170-1300) available in 155	476.7 (214-1084) available in 29	0.26
C-Reactive Protein (CRP)	6.2 (2-20.6) available in 208	5.1 (1.6-19.8) available in 177	12.6 (4.95-30.35) available in 31	0.01
Weighted Pediatric Crohn's Disease Activity Index (PCDAI)	25 (7.5-40) available in 198	23.75 (10-40.625) available in 172	26.25 (0-36.875) available in 26	0.19
GGA on 100mm VAS	32 (17-60)	31 (16-60)	40 (20-68)	0.1
GGA Likert 1- Complete deep remission**	27(12)	23(12)	4(11)	0.47
GGA Likert 2 - Clinical remission**	41(18)	36(19)	5(14)	0.27
GGA Likert 3- Mild disease activity	61(27)	51(26)	10(29)	0.39

GGA Likert 4 - Moderate disease activity	70(31)	58(30)	12(35)	0.3
GGA Likert 5 - Severe disease activity	24(10)	21(11)	3(9)	0.34
TI SES-CD	-	3 (0-6)	-	-
remission (<3)		94(50)		
mild (between 3 and 6)		50(26)		
moderate (>7)		45(24)		

Table supp2: Basic characteristics of the entire cohort. Mean ±standard deviation, median (interquartile range), and frequency (%) are displayed as appropriate.

GGA=gastroenterologist global assessment; VAS=visual analogue scale; TI SES-CD = terminal ileum simple endoscopic score of Crohn's disease; * according to Paris classification; ** clinically and also endoscopically (as much as is known).

	Subsample n=132	Intubation n=107	No intubation n=25	P
Male (%)	76 (58)	60 (57)	16 (63)	0.23
Disease duration, years	2.25 (0-4.4)	2.2 (0.15-4.1)	2.5 (0-4.8)	0.35
Age at enrollment, years	14.44±2.24	14.47±2.23	14.29±2.33	0.25
L1 (TI and cecum) *	40 (30)	32 (30)	8 (32)	0.42
L2 (Colonic)*	11(8)	9(8)	2(8)	0.47
L3 (Ileocolonic)*	80 (60)	65 (60)	15 (60)	0.47
L4 (Isolated upper GI)*	1(1)	1(1)	0(0)	0.32
Fecal Calprotectin (FC)	652 (226-1353)	657 (233-1475)	429.6 (224-1084)	0.09
	available in 106	available in 85	available in 21	
C-Reactive Protein (CRP)	8.5 (2.4-25.8)	7.3 (1.8-25.8)	12.65 (4.98-24.68)	0.14
	available in 124	available in 100	available in 24	
Weighted Pediatric Crohn's Disease Activity Index (PCDAI)	25 (10-40)	25 (10-40)	22.5 (0-37.5)	0.42
	available in 119	available in 98	available in 21	
PICMI	22.45±10.13	21.5±10.43	26.52±7.63	0.015
MaRIA	22.48±7.75	22.54±8	22.2±6.77	0.28
GGA VAS (0-100mm)	39 (20-60)	37 (20-60)	40 (22-70)	0.22
GGA Likert 1- Complete deep remission**	13(10)	11(10)	2(8)	0.36
GGA Likert 2 - Clinical remission**	19(14)	14(13)	5(20)	0.19

GGA Likert 3- Mild disease activity	39(30)	32(30)	7(28)	0.42
GGA Likert 4 - Moderate disease activity	47(35)	39(36)	8(32)	0.34
GGA Likert 5 - Severe disease activity	14(11)	11(10)	3(12)	0.4
TI SES-CD	-	5 (0-8)	-	-
remission (<3)		37(35)		
mild (between 3 and 6)		32(30)		
moderate (>7)		38(35)		

Table supp3: Basic characteristics of the group with all relevant MRE variables and Ileocolonoscopy. Mean +/- standard deviation, median (interquartile range), and frequency (%) are displayed as appropriate. GGA=gastroenterologist global assessment; VAS=visual analogue scale; TI SES-CD = terminal ileum simple endoscopic score of Crohn's disease; * according to Montreal classification; ** clinically and also endoscopically (as much as is known).

427

Parameter	Value
bootstrap	True
ccp_alpha	0
criterion	Mse
max_depth	2
max_features	Auto
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2

min_weight_fraction_leaf	0
n_estimators	100
n_jobs	None
oob_score	True
random_state	0
verbose	0
warm_start	False

Table supp4: Random forest algorithm parameters