Magnetic Resonance Enterography indices for ileal Crohn's disease assessment: Comparison between classical and machine-learning-based indices

Itai Guez, BSc[1], Gili Focht, MBA[2], Mary-Louise C. Greer, MBBS[3], Ruth Cytter-Kuint, MD[4], Li-tal Pratt, MD[5], Denise A. Castro, MD[5], Dan Turner, MD, PhD[2], Anne M. Griffiths, MD[6], Moti Freiman, PhD[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, MSc candidate, Technion - Israel Institute of Technology, Technion City, Haifa 3200003, Israel, +972-77-887-4147, itai.guez@campus.technion.ac.il

Abstract:

Rationale and Objectives: Magnetic Resonance Enterography (MRE) is a non-invasive examination to assess Crohn's Disease (CD) activity. Several MRE-based indices were developed for standardized evaluation. The goal of this study was to compare non-linear machine-learning models to linear models in evaluating terminal ileum (TI) inflammation severity from MRE data of pediatric CD patients.

Materials and Methods: We developed Random-Forest machine-learning (RF) and baseline multiple linear-regression (MLR) models to evaluate CD severity at the TI from MRE data with the Magnetic Resonance Index of Activity (MaRIA), it's simplified version (sMaRIA), and the Pediatric Inflammatory Crohn's MRE Index (PICMI) variables. We assessed the accuracy of the MRE-based TI CD severity evaluation with ileocolonoscopy-based Simple Endoscopic Score for CD (SES-CD) as the reference standard using a stratified 2-fold validation experimental setup, repeated 100 times.

Results: A total of 163 children who had all relevant MRE variables and available TI SES-CD score were included in this study (51% in remission [TI SEC-CD<3], 25% moderate [3≤ TI SEC-CD≤6], 24% high [TI SEC-CD ≥7]). RF-PICMI achieved the best R-squared coefficient on validation set compared to other models (Median value over 200 folds: RF-MaRIA 0.35 vs. MLR-MaRIA 0.29, RF-sMaRIA 0.37 vs. MLR-sMaRIA 0.27, and RF-PICMI 0.39 vs. MLR-PICMI 0.33, p<1e-5). Receiver operation characteristics (ROC) area under curve (AUC) for classification of mucosal healing [TI SEC-CD<3] was broadly similar in all models.

Conclusion: Machine-learning models with MRE variables as inputs might be more accurate than their linear-regression counterpart models in evaluating CD activity from MRE data.

Keywords: Crohn's disease, Pediatrics, Diagnostic performance, Magnetic resonance enterography, Machine learning.

[Type here]

38 **1. Introduction**

39 Treatment goals for Crohn's disease (CD) have shifted in the past two decades from symptom control towards
40 more objective goals in accordance to a "treat-to-target" philosophy.[1,2] Specifically, healing of inflammatory
41 lesions defined as endoscopic healing (EH)[3] is currently considered a primary objective treatment goal in
42 clinical trials.[4] The Simple Endoscopic Score for CD (SES-CD)[5] assessed by ileocolonoscopy is considered the
43 most established endoscopic score for EH evaluation.[6–8] A threshold value which is selected empirically by
44 experts or a reduction >50% from a baseline SES-CD score is considered as EH.[5]

45 During the past decade, Magnetic resonance enterography (MRE) has emerged as an imaging modality for
46 evaluating bowel healing.[3,9] MRE-based evaluation of healing in CD commonly takes the term "transmural
47 healing" (TH) as a synonym for complete healing of all the bowel layers (including the mucosa). Nevertheless,
48 there is still no validated definition of what constitutes TH in CD.[10] Current consensus for selecting therapeutic
49 targets in IBD (STRIDE-II) is that TH and histological healing are not formal targets but should be assessed as
50 measures of the remission depth, which means assessing healing from cross-sectional imaging is considered an
51 adjunct and not a treatment target.[11]
52 Several MRE-based indices were developed for standardized evaluation of TH from MRE data.[12] The most
53 common MRE index of CD activity in adults is the Magnetic Resonance Index of Activity (MaRIA) which includes
54 wall-thickness, ulcers, relative contrast enhancement (RCE) and wall edema.[13] Recently, a simplified version of
55 the MaRIA (sMaRIA) was introduced.[14] This version does not require RCE and thereby avoids intravenous
56 gadolinium-based contrast agents and includes fat stranding instead. More recently, the Pediatric
57 Inflammatory Crohn's MRE Index (PICMI) which also does not require RCE and avoids intravenous gadolinium-
58 based contrast agents and includes instead wall restricted diffusion, comb sign and mesenteric edema, has
59 been developed and validated specifically to assess TH in pediatric CD patients.[15]
60 Previous MRE index development and validation reference standards were primarily based on EH assessment

through correlation with endoscopic scores such as SES-CD[16] and Crohn's Disease Endoscopic Index of Severity (CDEIS)[17] as these indexes are considered reflective of the current medical gold standard while the TH definition has no consensus.[10]

Recently 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.[17–19] Yet, previous MRE-based indices development relied upon linear-regression models, assuming only a linear correlation between MRE items and the endoscopic score. Therefore, the goal of this study was to systematically compare diagnostic performances of machine-learning methods over classical methods for the development of MRE-based indices for ileal Crohn's disease assessment.

The presented models are available to the community through a dedicated website at: https://tcml-bme.github.io/ML_SESCD.html. Parts of this work were presented at the RSNA 2021 and the WSPGHAN 2021 conferences. A preliminary version of this work is available as a pre-print at https://doi.org/10.1101/2021.08.29.21262424

76 ## 2. Materials and Methods

77 ### 2.1 Ethical Approval

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

81 ### 2.2 Study population

82 This is a substudy of the multicenter ImageKids study (NCT01881490[20]) in which a total of 240 children who
83 fulfilled the eligibility criteria of being 6 to 18 years old and having an established diagnosis of CD enrolled to
84 the study. Enrolled participants underwent an ileocolonoscopy and MRE together with biosampling and
85 comprehensive collection of demographics, clinical and laboratory data within 14 days without any change in
86 treatment. Detailed description of the ImageKids study patient population and data collection is provided by
87 Focht et al.[15]

88 ImageKids study patients' data records with incomplete colonoscopies, prior bowel surgeries, non-intubated TI
89 during ileocolonoscopy, missing MRE information were excluded from our substudy.

90 ### 2.3 MRE protocol and radiological evaluation

91 The ImageKids MRE protocol included a localizer sequence and a motility sequence in the coronal plane,
92 followed by a series of coronal and/or axial rapid T2-weighted, T1-weighted gradient echo sequences and
93 diffusion-weighted imaging (DWI). T1-weighted sequences were performed pre- and post-intravenous
94 gadolinium-based contrast agent (GBCA) injection. An intravenous antispasmodic agent (glucagon or hyoscine
95 butylbromide) was administered following the motility sequence. The MRE was performed without an enema
96 as it is less feasible in pediatrics.

97 Each MRE was read twice by independent two independent central radiologists using an explicit case report
98 form who were chosen from a group of four pediatric radiologists: MLG, DC, RCK and LP, each with a minimum
99 of 10 years of experience, from four different sites, two in Canada and two in Israel). The radiologists were
100 blinded to the clinical and endoscopic data. Radiologists reported, among others, the MaRIA,[13] sMARIA,[14] and
101 the PICMI variables.[15] The TI was defined in the Imagekids study as the first 10cm segment of the ileum
102 measured from the ileocecal valve. Table supp1 in the supplementary materials summarizes the variables for
103 each index.

104 **2.4 Endoscopic report and gastroenterologist assessment**

105 The SES-CD evaluates four items (ulcer size, proportion of the surface area that is ulcerated, proportion of the
106 surface area affected, narrowing) by scoring each item on an ordinal scale from 0 to 3 in five bowel segments
107 (TI, right colon, transverse colon, left colon, rectum), where a higher score indicates a more severe disease
108 defined by a higher percentage of mucosal inflammation. The maximum score in a passable segment is 11
109 while a score of 12 can be obtained only in the last evaluated segment when there is impassable stenosis. The
110 total SES-CD score, which is the sum of the five segments scores, ranges from 0 to 56. No adjustment is made
111 to account for non-visualized segments (due to resection, technical difficulty, or stenosis).[5] The SES-CD items
112 were scored by a site gastroenterologist during ileocolonoscopy for each bowel segment. In addition,
113 Gastroenterologist global assessment (GGA) of the degree of inflammation was scored on a 100-mm visual
114 analogue score (VAS) and on a Likert scale from 1 (i.e. Complete deep remission) to 5 (i.e. Severe disease
115 activity). The site gastroenterologists were un-blinded to clinical data. For the purpose of this study only the TI
116 portion of the SES-CD score was used and not the rest of the score.

117 **2.5. Development of Machine-Learning-based MRE indices for CD**

118 We developed three non-linear machine-learning-based MRE indices for CD severity assessment using the
119 random-forest algorithm,[21] each using a different set of variables. The first index used the MaRIA variables (RF-

120 MaRIA), the second index used the sMaRIA variables (RF-sMaRIA), and the third index used the PICMI variables
121 (RF-PICMI). In addition, we constructed three multiple-linear-regression-based indices, each using the same
122 three sets of variables as input features (MLR-MaRIA, MLR-sMaRIA, MLR-PICMI). The linear regression models
123 produce a continuous score which is then rounded to nearest integer value, while the random-forest models
124 produce an ordinal score. The models' scores were normalized to a 0-1 range by dividing it by the maximum
125 possible SES-CD value of 12.

126 The TI SES-CD score was used as the reference for indices' development and validation. Therefore, we used
127 only patients with an available TI SES-CD score which had all MRE variables of both indices. Table supp2
128 summarizes the full parameter list used for configuration and optimization of the machine-learning algorithms.
129 The models were written in Python [version 3.6.4] using the open-source Scipy [version 1.5.4] library and the
130 open-source Statsmodels [version 0.11.0] library. Models ran on an Intel i3 CPU.

131 **2.5 Statistics**

132 We compared diagnostic performance of the machine-learning-based MRE indices to this of their counterpart
133 multiple-linear-regression-based indices using the TI SES-CD as the reference-standard. We defined distinct
134 groups of patients for development (50%) and validation (50%). We mitigated the bias of the selection of
135 patients to development and validation sets and properly compare the set of variables and the chosen
136 algorithm by using stratified 2-fold validation approach repeated 100 times, with a total of 200 folds.

137 We assessed indices accuracy of the different models by means of the R-squared coefficient with the reference
138 endoscopic TI SES-CD. We determined which set of variables (PICMI, MaRIA, or sMARIA) provides the most
139 accurate predictions in comparison to the reference endoscopic TI SES-CD measurements by means of the
140 highest R-squared. We used the non-parametric Wilcoxon non-parametric test with Bonferroni correction to
141 control the family wise error rate (FWER) with $p<0.01$ to assess the statistical significance of the differences in
142 the R-squared. Finally, we evaluated the capacity of the different indices to distinguish between patients with

143 and without MH in comparison with reference MH assessments defined as reference SES-CD<3,[4,22] using the

144 area under the receiver operating characteristic (ROC) curves (AUC).

145

146

147 ## 3. Results

148 ### 3.1 Study population

149 Figure 1 summarizes the patients selection process for the study. Of the 240 children included in the original
150 ImageKids study, a total of 163 children were included in this sub-study, while 77 were excluded (5 due to
151 incomplete colonoscopies, 13 due to prior bowel surgeries, 33 due to non-intubated TI and 26 due to missing
152 one or more MRE items [wall restricted diffusion (n=16), RCE (n=22)]).

153 Of the 163 children included in this sub-study, 51% were in remission [TI SES-CD<3], 25% moderate [3≤TI SES-
154 CD≤6], 24% high [TI SES-CD ≥7]). 107 out of the 163 children (65%) had some degree of inflammation according
155 to radiologists reports while 56 (35%) did not have any signs of inflammation seen in the TI based on MRE (all
156 MRE features were zeroes). only 6 of the 56 without any MRE signs of inflammation had a TI SES-CD>3.
157 Table supp3 summarizes the demographic and variables values for the patients included in this sub-study.

158 ### 4.2 TI SES-CD prediction from MRE

159 The stratified 2-fold validation evaluation scheme divided the set of 163 patients into 81/82 for
160 development/validation 200 times. Table 1 summarizes the distribution of the R-squared between the models'
161 CD severity assessment and reference TI SES-CD for the different models on the validation sets. The machine-
162 learning-based models achieved a better R-squared compared to their multiple-linear-regression counterparts
163 for each set of variables. The PICMI-based models achieved a better R-squared over MARIA/sMARIA-based
164 models for both the machine-learning-based models and the multiple-linear-regression models. Fig. 2 depicts
165 the R-squared distribution of the different validation sets. All differences were statistically significant.

166

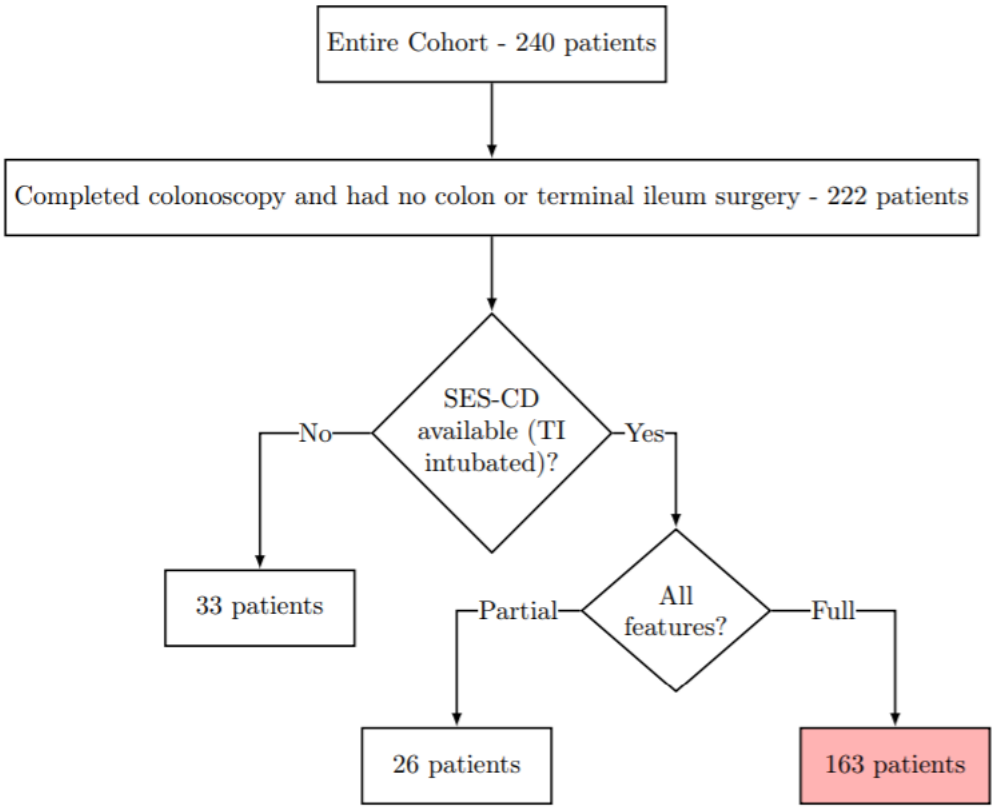

Figure 1: Patients selection from the entire cohort. MRE: Magnetic Resonance Enterography

R-squared distribution	Median	Q1	Q3
MLR-MaRIA	0.29	0.22	0.33
MLR-sMaRIA	0.27	0.22	0.33
MLR-PICMI	0.33	0.26	0.38
RF-MaRIA	0.35	0.29	0.4
RF-sMaRIA	0.37	0.3	0.41
RF-PICMI	**0.39**	**0.33**	**0.42**

Table 1: Validation set R-squared distribution statistics over the folds per model, RF: random-forest, MLR: multiple-linear-regression. Best result in bold.

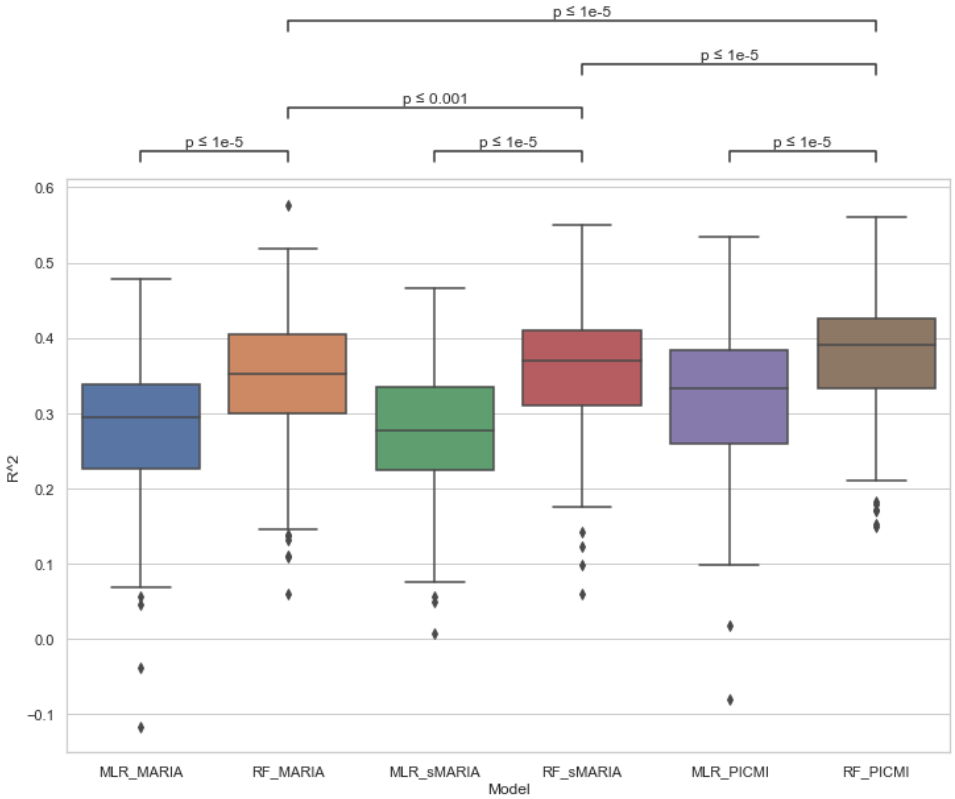

Figure 2: R-squared distribution over the validation folds

Fig. 3 presents the averaged ROC curves of the different models for the differentiation between patients who achieved MH (TI SES-CD<3) and patients who did not achieve MH (TI SES-CD≥3) according to the reference ileocolonscopy-based TI SES-CD score. The RF-PICMI achieved the best distribution of AUC over the different folds. However, the improvement was not statistically significant.

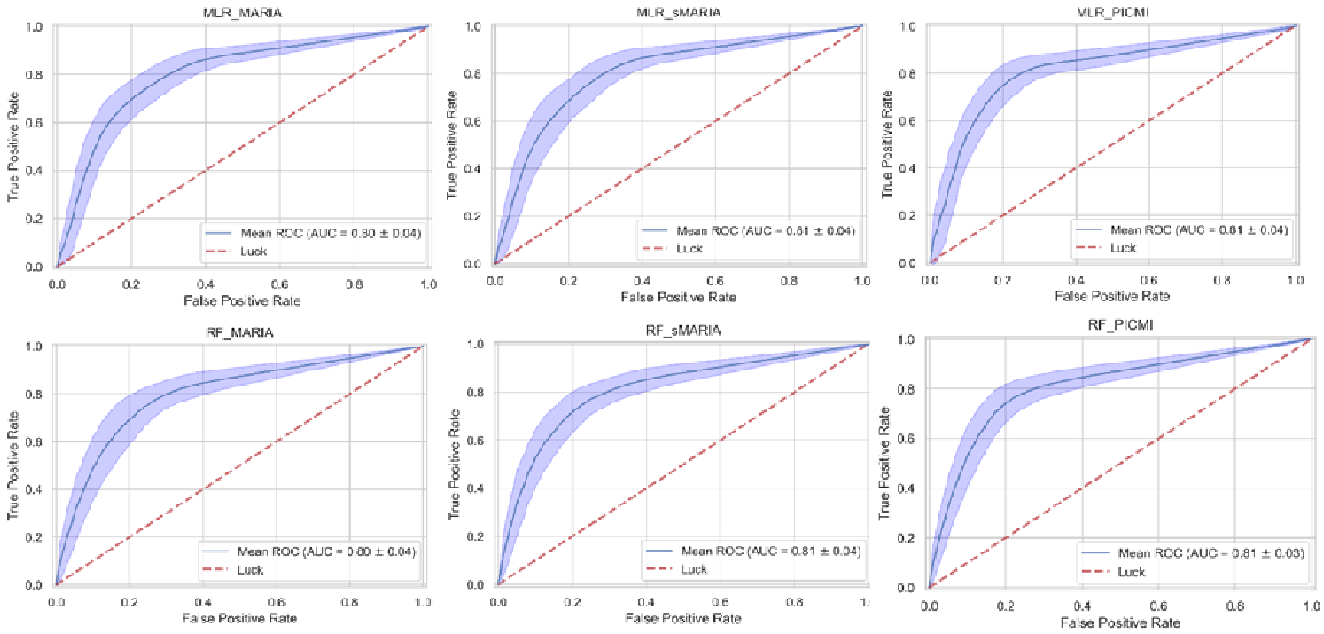

Figure 3: Average ROC curves over the different folds for all models.

4. Discussion

The development of MRE biomarkers, MRE indices and the definition of TH play an important role in objective assessment of CD severity with MRE.[23] Yet, the reference standard for evaluating MRE indices for CD severity assessment remains endoscopic scores such as the SES-CD.[24] However, current MRE indices achieved only a moderate correlation with the reference endoscopic scores. In this study, we hypothesized that the moderate correlation to the TI SES-CD achieved by MRE indices may be attributed, at least in part, to the utilization of classical linear regression models to characterize the correlation between the MRE-based variables and EH which limited their ability to determine the potentially complex, non-linear correlation.[25]

Our results demonstrated that non-linear machine-learning models were better in characterizing the correlation between the MRE items and EH compared to their linear counterparts. The machine-learning models improved the correlation between the CD severity assessment by each MRE-based index (MaRIA,

193 sMaRIA and PICMI) and the reference TI SES-CD. The combination the PICMI MRE variables and the
194 machine-learning-based model yielded the highest correlation between the model's CD severity
195 assessment and the reference TI SES-CD.

196 The better performance of the machine-learning-based models compared to their multiple- linear-
197 regression counterpart was expected as the machine-learning-based models have a better capacity to
198 characterize complex, non-linear correlations between input variables and outcomes compared to linear
199 models.[25] Specifically, we chose to use the RF model[21] for the machine-learning models development as it
200 is a well-established model useful for small datasets and structured data. The RF ensemble mechanism
201 controls overfitting by built-in feature selection and especially suits discrete target as it had a fixed set of
202 possible values covered by the dataset.

203 Yet, the improved correlation between the models' predictions and the reference TI SES-CD did not lead to
204 an improvement in classification between patients who achieved MH and patients who did not achieve
205 MH by means of the AUC. This may be explained as a result of improved correlation demonstrated mostly
206 on patients with SES-CD higher than 2 rather than in the entire range of scores due to limited availability of
207 data. We expect that training the models on larger amount of data will result in improved classification
208 accuracy as well.

209 While, to the best of our knowledge, we are the first to propose using non-linear machine-learning models
210 for the development of MRE-based indices for CD severity assessment, the utilization of machine-learning
211 models in the domain of IBD is not new. Specifically, Dhaliwal et al. used a machine-learning model to
212 distinguish pediatric colonic IBD subtypes based on clinical, endoscopic, radiologic, and histologic data[26]
213 and Waljee et al. demonstrated the added value of a machine-learning model to predict disease course
214 from demographic data and clinical records compared to a linear-regression model.[27] Additional works
215 aimed to predict endoscopic scores of CD severity from ilecolonoscopy data.[28,29]

In addition, our results demonstrated an improved agreement between the MRE-based assessment and reference TI SES-CD when using the PICMI variables over the MaRIA and sMaRIA variables in the model. Although previous literature questioned the need to develop a specific MRE index of activity for the pediatric population,[3] the improved agreement achieved by the PICMI based models suggest an added-value to MRE-indices for CD severity assessment developed specifically to the pediatric population.

Finally, our best-performing model (RF-PICMI) used a DWI sequence to assess wall restricted diffusion variable rather than a post-gadolinium T1-weighted sequence. This finding suggests that the diagnostic accuracy of DWI, included in PICMI and not MaRIA and sMaRIA, is not overly estimated as was previously hypothesized.[14] Further, it provides some evidence that the post-contrast sequence can be omitted on MRE in children to reduce use of GBCA although it is not yet standard practice.[30]

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 readings. 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 is assumed to require more data compared to the development of linear models due to their complexity, we expect better results when the models are trained on the entire dataset, however then we cannot define properly held out validation set. Therefore, the presented models' performances might underestimate the added-value of the machine-learning models.

A future study to completly reproduce our results would require external validation by means of a fully independent test dataset.

239 Finally, CD is a transmural disease rather than a mucosal disease. It is not clear yet whether mucosal or
240 transmural healing should be the primary treatment goal.[10] While MRE has the ability to detect CD lesions
241 beyond the mucosa, its ability to assess low severity mucosal lesions is much more limited.[3] Therefore,
242 combined assessment with MRE and ileocolonscopy is ideal to fully assess the extent of the disease.

243 **Conclusions**

244 Our study demonstrates the important role of the combination of non-linear machine-learning models and
245 the selection of specific MRE-based variables to improve our ability to assess the severity of CD from MRE
246 data. Such machine-learning methods have the potential to improve MRE-based IBD assessment in
247 additional applications.
248 PICMI-based models had better results than MaRIA/sMaRIA-based models using both the random-forest
249 and multiple-linear-regression models. Further, our evaluation approach can serve as a research tool for
250 the clinical community to determine the most predictive features in clinical challenges that are still lacking
251 consensus about the most appropriate data acquisition, and to guide more efficient data collection in large
252 prospective studies of machine-learning applications in healthcare. The presented models to CD severity
253 from MRE-based variables are available to the community through a dedicated website at: https://tcml-
254 bme.github.io/ML_SESCD.html.

255

256

257

258

259 ## 5. References

260 1. Peyrin-Biroulet L, Sandborn W, Sands BE, et al. Selecting Therapeutic Targets in Inflammatory Bowel
261 Disease (STRIDE): Determining Therapeutic Goals for Treat-to-Target. *Am J Gastroenterol*.
262 2015;110(9):1324-1338. doi:10.1038/ajg.2015.233

263 2. Colombel JF, D'Haens G, Lee WJ, Petersson J, Panaccione R. Outcomes and Strategies to Support a
264 Treat-to-target Approach in Inflammatory Bowel Disease: A Systematic Review. *J Crohn's Colitis*.
265 2020;14(2):254-266. doi:10.1093/ecco-jcc/jjz131

266 3. Panés J, Rimola J. Is the Objective of Treatment for Crohn's Disease Mucosal or Transmural Healing?
267 *Clin Gastroenterol Hepatol*. 2018;16(7):1037-1039. doi:10.1016/j.cgh.2018.03.034

268 4. Turner D, Griffiths AM, Wilson D, et al. Designing clinical trials in paediatric inflammatory bowel
269 diseases: a PIBDnet commentary. *Gut*. 2020;69(1):32-41. doi:10.1136/gutjnl-2018-317987

270 5. Daperno M, D'Haens G, Van Assche G, et al. Development and validation of a new, simplified
271 endoscopic activity score for Crohn's disease: The SES-CD. *Gastrointest Endosc*. 2004;60(4):505-512.
272 doi:10.1016/S0016-5107(04)01878-4

273 6. Dubcenco E, Zou G, Stitt L, et al. Effect of standardised scoring conventions on inter-rater reliability in
274 the endoscopic evaluation of Crohn's disease. *J Crohn's Colitis*. 2016;10(9):1006-1014.
275 doi:10.1093/ecco-jcc/jjw120

276 7. Khanna R, Bouguen G, Feagan BG, et al. A systematic review of measurement of endoscopic disease
277 activity and mucosal healing in Crohn's disease: Recommendations for clinical trial design. *Inflamm*
278 *Bowel Dis*. 2014;20(10):1850-1861. doi:10.1097/MIB.0000000000000131

279 8. Khanna R, Zou G, D'Haens G, et al. Reliability among central readers in the evaluation of endoscopic
280 findings from patients with Crohn's disease. *Gut*. 2016;65(7):1119-1125. doi:10.1136/gutjnl-2014-
281 308973

282 9. Walsh AJ, Bryant R V., Travis SPLL. Current best practice for disease activity assessment in IBD. *Nat Rev*
283 *Gastroenterol Hepatol*. 2016;13(10):567-579. doi:10.1038/nrgastro.2016.128

284 10. Serban ED. Treat-to-target in Crohn's disease: Will transmural healing become a therapeutic endpoint?
285 *World J Clin Cases*. 2018;6(12):501-513. doi:10.12998/WJCC.V6.I12.501

286 11. Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: An Update on the Selecting Therapeutic Targets in
287 Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD
288 (IOIBD): Determining Therapeutic Goals for Treat-to-Target strategies in IBD. *Gastroenterology*.
289 2021;160(5):1570-1583. doi:10.1053/j.gastro.2020.12.031

290 12. Rozendorn N, Amitai MM, Eliakim RA, Kopylov U, Klang E. A review of magnetic resonance
291 enterography-based indices for quantification of Crohn's disease inflammation. *Therap Adv*
292 *Gastroenterol*. 2018;11(1):1-21. doi:10.1177/1756284818765956

293 13. Rimola J, Ordás I, Rodriguez S, et al. Magnetic resonance imaging for evaluation of Crohn's disease:
294 Validation of parameters of severity and quantitative index of activity. *Inflamm Bowel Dis*.
295 2011;17(8):1759-1768. doi:10.1002/ibd.21551

296 14. Ordás I, Rimola J, Alfaro I, et al. Development and Validation of a Simplified Magnetic Resonance Index
297 of Activity for Crohn's Disease. *Gastroenterology*. 2019;157(2):432--439.e1.
298 doi:10.1053/j.gastro.2019.03.051

299 15. Focht G, Kuint RC, Greer M-LC, et al. Development, validation and evaluation of the Pediatric
300 Inflammatory Crohn's Magnetic Resonance Enterography Index (PICMI) from the ImageKids study.

301 *Gastroenterology*. 2022;0(0). doi:10.1053/J.GASTRO.2022.07.048

302 16. Zheng X, Li M, Wu Y, et al. Assessment of pediatric Crohn's disease activity: validation of the magnetic
303 resonance enterography global score (MEGS) against endoscopic activity score (SES-CD). *Abdom Radiol*.
304 2020;45(11):3653-3661. doi:10.1007/s00261-020-02590-8

305 17. Olivera P, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: a look into the future. *Nat Rev*
306 *Gastroenterol Hepatol*. 2019;16(5):312-321. doi:10.1038/s41575-019-0102-5

307 18. Le Berre C, Sandborn WJ, Aridhi S, et al. Application of Artificial Intelligence to Gastroenterology and
308 Hepatology. *Gastroenterology*. 2020;158(1):76-94.e2. doi:10.1053/j.gastro.2019.08.058

309 19. Seyed Tabib NS, Madgwick M, Sudhakar P, Verstockt B, Korcsmaros T, Vermeire S. Big data in IBD: Big
310 progress for clinical practice. *Gut*. 2020;69(8):1520-1532. doi:10.1136/gutjnl-2019-320065

311 20. https://clinicaltrials.gov/ct2/show/study/NCT01881490.

312 21. Breiman L. Random Forests. *Mach Learn*. 2001;45:5-32.

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

316 23. Bruining DH, Zimmermann EM, Loftus E V., et al. Consensus recommendations for evaluation,
317 interpretation, and utilization of computed tomography and magnetic resonance enterography in
318 patients with small bowel Crohn's disease. *Radiology*. 2018;286(3):776-799.
319 doi:10.1148/radiol.2018171737

320 24. Roseira J, Ventosa AR, de Sousa HT, Brito J. The new simplified MARIA score applies beyond clinical
321 trials: Asuitable clinical practice tool for Crohn's disease that parallels a simpleendoscopic index and

322 fecal calprotectin. *United Eur Gastroenterol J*. 2020;8(10):1208. doi:10.1177/2050640620943089

323 25. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: From Theory to Algorithms*. Vol
324 9781107057135. Cambridge University Press; 2013. doi:10.1017/CBO9781107298019

325 26. Dhaliwal J, Erdman L, Drysdale E, et al. Accurate Classification of Pediatric Colonic Inflammatory Bowel
326 Disease Subtype Using a Random Forest Machine Learning Classifier. *J Pediatr Gastroenterol Nutr*.
327 2021;72(2):262-269. doi:10.1097/MPG.0000000000002956

328 27. Waljee AK, Lipson R, Wiitala WL, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in
329 Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm Bowel Dis*. 2018;24(1):45-53.
330 doi:10.1093/ibd/izx007

331 28. Klang E, Grinman A, Soffer S, et al. Automated Detection of Crohn's Disease Intestinal Strictures on
332 Capsule Endoscopy Images Using Deep Neural Networks. *J Crohn's Colitis*. 2021;15(5):749-756.
333 doi:10.1093/ecco-jcc/jjaa234

334 29. Klang E, Barash Y, Margalit RY, et al. Deep learning algorithms for automated detection of Crohn's
335 disease ulcers by video capsule endoscopy. *Gastrointest Endosc*. 2020;91(3):606-613.e2.
336 doi:10.1016/j.gie.2019.11.012

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

339

340

341

342

6. Supplementary material

Variable	PICMI	MaRIA	sMARIA	Range	Definition
Wall-thickness	V	V	V	Positive Number (mm)	Measure the thickest part within the TI segment
Ulcers	V	V	V	No - 0, Yes - 1	Mucosal indentation smaller than the full thickn of the bowel wall
Perienteric fat stranding (assessed by mesenteric T2 hyperintensity)	V	-	V	No - 0, Yes - 1	T2 hyperintense signal of the mesentery adjacer to the same bowel loop assessed
Mural edema	-	V	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, judgec 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.
Wall restricted diffusion	V	-	-	0 - None/Mild, 1 – Moderate/Marked	A subjective measure of hyperintensity on highe b-value trace image compared to normal bowel
Comb sign	V	-	-	No - 0, Yes - 1	Vascular engorgement related to the bowel loop assessed

Table supp1: Variables of the MRE-based indices

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 supp2: Random forest algorithm parameters

369

	Subsample n=163	TI remission(MH) TI SES-CD < 3 n=83	No TI remission TI SES-CD >= 3 n=80	P
Male (%)	88 (54)	44 (53)	44 (55)	-
Disease duration, years	2.2 (0.45-4)	2.2 (1.1-3.95)	2.5 (0-4.8)	0.46
Age at enrollment, years	14.24±2.44	14.4±2.45	14.4±2.42	0.22
L1 (TI and cecum) *	37 (22)	13 (16)	24 (30)	-
L2 (Colonic)*	26(16)	23(27)	3(4)	-
L3 (Ileocolonic)*	96 (59)	43 (52)	53 (66)	-
L4 (Isolated upper GI)*	4(2)	4(5)	0(0)	-
Fecal Calprotectin (FC)	629 (151-1319) available in 132	393(70-806) available in 70	828 (532-1601) available in 62	7.9e-5
C-Reactive Protein (CRP)	4.95 (1.4-20.13) available in 152	4 (1-8.45) available in 75	9.5 (2.9-30.6) available in 77	0.003
Weighted Pediatric Crohn's Disease Activity Index (PCDAI)	22.5(10-40) available in 150	17.5 (0-37) available in 74	26.25 (15-41) available in 76	0.07
PICMI	21.5±10.43	15.56±10.46	24.64±9.02	2.7e-5
MaRIA	22.55±8	17.64±7.61	25.14±6.93	7.3e-6
GGA VAS (0-100mm)	31 (16-59)	19 (5-53)	40 (23-60)	6.1e-5
GGA Likert 1- Complete deep remission**	19(12)	19(23)	0(0)	-

GGA Likert 2 - Clinical remission**	30(18)	16(20)	14(18)	-
GGA Likert 3- Mild disease activity	46(28)	21(25)	25(31)	-
GGA Likert 4 - Moderate disease activity	50(31)	19(23)	31(39)	-
GGA Likert 5 - Severe disease activity	18(11)	8(10)	10(12)	-

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

375

376

377