Published in final edited form as:

Clin Gastroenterol Hepatol. 2018 August ; 16(8): 1354–1355.e1. doi:10.1016/j.cgh.2017.12.022.

Machine Learning in Detection of Undiagnosed Celiac Disease

Isabel A. Hujoel, M.D.¹, Dennis H. Murphree Jr, Ph.D.², Carol T. Van Dyke³, Rok Seon Choung, M.D., Ph.D.³, Ayush Sharma, M.B.B.S³, Joseph A. Murray, M.D.³, and Alberto Rubio-Tapia, M.D.³

¹Division of Internal Medicine, Mayo Clinic, Rochester, MN 55905

²Division of Health Sciences Research, Mayo Clinic, Rochester, MN 55905

³Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN 55905

Introduction

Celiac disease (CD) has a diversity of presentations that include typical symptoms such as diarrhea and weight loss and atypical symptoms such as elevated liver function tests and fatigue. Current guidelines recommend case-finding for diagnosis of CD, a technique in which physicians test individuals thought to be at risk for the disease due to the presence of conditions or symptoms associated with CD¹. However CD, particularly cases with atypical symptoms, remains largely undiagnosed^{2,3}. Increasing evidence suggests that case-finding may not be effective^{4,5}. Given the morbidity of undiagnosed symptomatic CD, a new method of detection needs to be developed⁶. We therefore sought to build a machine learning model for identifying undiagnosed CD based on permutations of currently accepted indications to test for CD.

Methods

Serum samples were collected from 47,557 individuals without prior CD diagnosis, receiving care at Mayo Clinic (Rochester, Minnesota) between 1995 and 2009. 408 undiagnosed cases were identified through sequential serology: a positive/indeterminate immunoglobulin A tissue transglutaminase antibody followed by a positive endomysial antibody. 408 gender, age, and year of serum collection -matched controls were selected

Corresponding Author: Alberto Rubio-Tapia, M.D., Division of Gastroenterology and Hepatology, Mayo Clinic, 200 First Street S.W., Rochester, Minnesota 55905, (507)284-2511, rubiotapia.alberto@mayo.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosures: None; all authors have nothing to disclose

Specific author contributions:

IH: concept, drafted manuscript, data acquisition, interpretation of data

DM: statistical analysis, contributed to manuscript preparation

CVD: identified cases of CD

RSC: contributed to manuscript preparation

AS: data acquisition, contributed to manuscript preparation

JAM: Study design, input on content and analysis, contributed to manuscript preparation, financial support

ART: concept, study design, contributed to manuscript preparation, interpretation of data

from individuals with negative composite serology. Medical records were reviewed by a physician blinded to serologic status. Using a standardized collection sheet, they collected data on frequency of indications to test for CD drawn from current guidelines (Supplementary Table)¹.

To identify undiagnosed cases we built a series of predictive models utilizing a variety of statistical approaches: logistic regression, elastic net, tree-based models with and without boosting and/or bagging, support vector machine with radial basis functions, a neural network (single layer perceptron), random forest, and linear discriminant analysis. All models were built using a grid search over plausible hyperparameters, with performance assessed via area under the receiver-operator curve (AUC) on out-of-sample observations during ten-fold cross-validation. Pre-processing consisted of the centering and scaling of all continuous variables, the removal of all zero-variance predictors, and the omission of any records with missing variables (20 observations, 2.5% of our cohort).

Experiments were conducted both including and excluding variables with low variance. This meaningfully affected the performance of only the elastic net, which is unsurprising given its regularization. Due to this minor effect, low and zero variance variables were removed, and all results considered here are for the predictor set: sex, age, number of symptoms, history of any autoimmune condition, thyroid disorder, anemia, hypothyroidism, previous indication to test for CD, dyspepsia, and recurring abdominal pain.

Results

Of the ten models trained, all yielded poor discriminatory performance, with AUC's ranging from 0.49 to 0.53 (Figure 1). Only two, random forest and bagged classification trees, outperformed random chance with a likelihood greater than 95%, and these were not significantly different from the other models.

Discussion

Only two models outperformed random chance in identifying undiagnosed CD and these had very low predictive power. The inability to develop a successful model suggests that currently used indications to test do not characterize undiagnosed CD. This has significant clinical implications for how providers should screen for CD, which remains largely undetected.

Atypical and asymptomatic cases make up the majority of the estimated 80% of undiagnosed CD². While evidence suggests that atypical cases benefit from diagnosis, the US Preventive Services Task Force recently determined there was insufficient evidence to support screening asymptomatic individuals^{7,11, 15}. Case-finding has therefore been proposed to increase detection of atypical cases. However, our findings, along with several recent studies, suggest that case-finding is not as effective, and that continued use may not detect most cases with CD^{1,4,5}.

Our failure to develop an effective model may be due to the subtle symptoms in atypical cases and suggests that case-finding based on symptoms may be ineffective. Prior studies

found that people detected through screening had mild symptoms that went unrecognized until they improved with treatment⁸. If the nature of undiagnosed cases is that symptoms do not come to clinical attention, then case-finding strategies looking for symptoms will not be effective. Conversely, strategies potentially looking at conditions associated with CD or physical exam findings may be more effective. Our study was limited in that it could not accurately evaluate family history, which has previously been associated with a high risk of CD.

In conclusion, we were unable to develop an effective model based on currently used indications to test to identify undiagnosed CD cases. This suggests providers cannot rely on currently accepted symptomatology to determine whom to screen. Case-finding based on these symptoms is unlikely to be effective in increasing diagnosis, and new methods need to be explored.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: NIH RO1-DK57892 (JAM)

Abbreviations

CD Celiac disease

References

- 1Rubio-Tapia A, Hill ID, Kelly CP, Calderwood AH, Murray JA. American College of Gastroenterology Clinical Guideline: Diagnosis and Management of Celiac Disease. *American Journal of Gastroenterology*. 2013; 108:656–677. [PubMed: 23609613]
- 2Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, Everhart JE. The Prevalence of Celiac Disease in the United States. *The American Journal of Gastroenterology*. 2012; 107:1538–1544. [PubMed: 22850429]
- 3Choung RS, Unalp-Arida A, Ruhl CE, Brantner TL, Everhart JE, Murray JA. Less Hidden Celiac Disease But Increased Gluten Avoidance Without a Diagnosis in the United States: Findings from the National Health and Nutrition Examination Surveys From 2009 to 2014. *Mayo Clinic Proceedings*. 2017; 92:30–38.
- 4Rosén A, Sandström O, Carlsson A, et al. Usefulness of Symptoms to Screen for Celiac Disease. *Pediatrics*. 2014; 133:211–218. [PubMed: 24420802]
- 5van der Windt DAWM, Jellema P, Mulder CJ, Kneepkens CMF, van der Horst HE. Diagnostic Testing for Celiac Disease Among Patients with Abdominal Symptoms: A Systematic Review. *JAMA*. 2010; 303:1738–1746. [PubMed: 20442390]
- 6Choung RS, Larson SA, Khaleghi S, et al. Prevalence and Morbidity of Undiagnosed Celiac Disease From a Community-Based Study. *Gastroenterology*. 2017; 152:830–839. [PubMed: 27916669]
- 7Bibbins-Domingo K, Grossman D, Curry S, et al. Screening for Celiac Disease: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2017; 317:1252–1257. [PubMed: 28350936]
- 8Kivelä L, Kaukinen K, Huhtala H, Lähdeaho M-L, Mäki M, Kurppa K. At-Risk Screened Children with Celiac Disease are Comparable in Disease Severity and Dietary Adherence to Those Found

because of Clinical Suspicion: A Large Cohort Study. *The Journal of Pediatrics*. 2017; 183:115–121. [PubMed: 28153477]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Area under the receiver-operator curve (AUC) from out-of-sample cross-validation. Only the bagged tree (treebag) and random forest (rf) models outperform random chance.