THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

CD or not CD, that is the question - a digital interobserver agreement study in coeliac disease

Citation for published version:
Denholm, J, Schreiber, BA, Jaeckle, F, Wicks, M, Benbow, EW, Bracey, TS, Chan, JYH, Farkas, L, Fryer, E, Gopalakrishnan, K, Hughes, C, Kirkwood, KJ, Langman, G, Mahler-Araujo, B, McMahon , RFT, La Win Myint, K, Natu, S, Robinson, A, Sanduka, A, Sheppard, KA, Tsang, YW, Arends, MJ & Soilleux, EJ 2024, 'CD or not CD, that is the question - a digital interobserver agreement study in coeliac disease', *BMJ open gastroenterology*. https://doi.org/10.1136/ bmjgast-2023-001252

Digital Object Identifier (DOI):
10.1136/ bmjgast-2023-001252

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
BMJ open gastroenterology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 11. Feb. 2024

BMJ
Open
Gastroenterology

CD, or not CD, that is the question: a digital interobserver agreement study in coeliac disease

James Denholm,[1,2,3] Benjamin A Schreiber,[1,2] Florian Jaeckle ,[1,3] Mike N Wicks,[4] Emyr W Benbow,[5,6] Tim S Bracey,[7,8] James Y H Chan,[9] Lorant Farkas,[10,11] Eve Fryer,[12] Kishore Gopalakrishnan,[13] Caroline A Hughes,[12] Kathryn J Kirkwood,[14] Gerald Langman,[15] Betania Mahler-Araujo ,[9,16] Raymond F T McMahon,[5,6] Khun La Win Myint,[17] Sonali Natu,[18] Andrew Robinson,[13] Ashraf Sanduka,[9] Katharine A Sheppard,[12] Yee Wah Tsang,[13] Mark J Arends,[19] Elizabeth J Soilleux[1,3]

To cite: Denholm J, Schreiber BA, Jaeckle F, *et al*. CD, or not CD, that is the question: a digital interobserver agreement study in coeliac disease. *BMJ Open Gastroenterol* 2024;**11**:e001252. doi:10.1136/bmjgast-2023-001252

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjgast-2023-001252).

Received 12 September 2023
Accepted 11 December 2023

Check for updates

© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Florian Jaeckle;
florian.jaeckle@gmail.com

ABSTRACT

Objective Coeliac disease (CD) diagnosis generally depends on histological examination of duodenal biopsies. We present the first study analysing the concordance in examination of duodenal biopsies using digitised whole-slide images (WSIs). We further investigate whether the inclusion of immunoglobulin A tissue transglutaminase (IgA tTG) and haemoglobin (Hb) data improves the interobserver agreement of diagnosis.

Design We undertook a large study of the concordance in histological examination of duodenal biopsies using digitised WSIs in an entirely virtual reporting setting. Our study was organised in two phases: in phase 1, 13 pathologists independently classified 100 duodenal biopsies (40 normal; 40 CD; 20 indeterminate enteropathy) in the absence of any clinical or laboratory data. In phase 2, the same pathologists examined the (re-anonymised) WSIs with the inclusion of IgA tTG and Hb data.

Results We found the mean probability of two observers agreeing in the absence of additional data to be 0.73 (±0.08) with a corresponding Cohen's kappa of 0.59 (±0.11). We further showed that the inclusion of additional data increased the concordance to 0.80 (±0.06) with a Cohen's kappa coefficient of 0.67 (±0.09).

Conclusion We showed that the addition of serological data significantly improves the quality of CD diagnosis. However, the limited interobserver agreement in CD diagnosis using digitised WSIs, even after the inclusion of IgA tTG and Hb data, indicates the importance of interpreting duodenal biopsy in the appropriate clinical context. It further highlights the unmet need for an objective means of reproducible duodenal biopsy diagnosis, such as the automated analysis of WSIs using artificial intelligence.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Concordance studies in coeliac disease (CD) diagnosis using glass slides have shown low levels of agreement between pathologists. The observed agreement varies from κ=0.3 to due to the general lack of κ=0.9 standardisation in the studies' designs and the small number of different pathologists participating in most of the existing work.

WHAT THIS STUDY ADDS

⇒ This first-in-class, large-scale concordance study of the histological diagnosis of CD based on digital whole-slide images gave a general concordance for histological diagnosis of CD of 0.73 (±0.08). Including additional data (IgA tTG and Hb) improved the agreement by 10%.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study shows that pathologist concordance in diagnosing CD using digital images is low. It is slightly improved by access to serological data and haemoglobin level. It provides a clear rationale for the development of a more reproducible and objective approach to the assessment of duodenal biopsies. Such a new approach could include the use of artificial intelligence. However, in that case this study also highlights the need to develop very carefully curated datasets, in which diagnostic accuracy (ground truth) is optimised, ideally by considering histopathology, serological results, haemoglobin level and clinical data together.

INTRODUCTION

In the autoimmune enteropathy, coeliac disease (CD), the ingestion of gluten (proteins found in wheat, barley and rye) results in a spectrum of relatively stereotyped changes in the duodenum.[1–7] The global prevalence of CD is ≈1%, while the global prevalence of biopsy-confirmed CD varies between 0.4% and 0.5% in South America and Africa, 0.4% and 0.8% in Europe, with a 0.6% prevalence in North America and Asia.[8] The prevalence is particularly high in the Celtic nations: CD-related hospital admissions in Scotland and Ireland have been reported two and three times as high as those in England,[9] and

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

the prevalence in Northern Ireland has been reported as high as ≈6%.[10] In countries such as Denmark and Scotland, the incidence is increasing.[11 12]

CD diagnosis in adults is generally based on histological examination of duodenal biopsies, which are preceded by measurement of immunoglobulin A tissue transglutaminase (IgA tTG)—and often endomysial antibody (EMA)—levels, and the consideration of clinical symptoms. While there is no definitive standard for diagnosing CD, some schemes have been suggested.[13–15]

The National Institute for Health and Care Excellence (NICE) guidelines which serve as 'evidence-based recommendations for health and care in England and Wales', suggest serological testing for those with CD-related symptoms, or first-degree relatives with a diagnosis, before referral to a gastrointestinal (GI) specialist for a biopsy.[16] The NICE guidelines also suggest a biopsy in cases where the serology is negative, but the symptoms persist.[16] The British Society of Gastroenterology (BSG) concluded biopsy remains essential for adult diagnosis,[17] although during the COVID-19 pandemic, the BSG recommended treating patients younger than 55 years old with suspected CD on the basis of IgA tTG serology alone.

Duodenal biopsies show a spectrum of histopathological appearances between CD and normal (as well as those of other rarer pathologies), rendering definitive diagnosis less straightforward than one might expect. Formal attempts to standardise histological examination of duodenal biopsies in the context of CD are used primarily in research/clinical trials and include the Marsh–Oberhuber scheme,[18 19] the Corazza–Villanacci scheme[20] and Ensari's method.[21]

Studies of the interobserver agreement in histological CD diagnosis are difficult to compare due to inconsistencies in their design: some included serology, while others have used histology alone. In some, the Marsh–Oberhuber or the Corazza–Villanacci schemes have been used, while others simply attempted binary classification of biopsies.

Arguelles-Grande *et al*[22] compared the agreement between a single pathologist and existing diagnoses, using the full Marsh–Oberhuber scale, on 102 biopsies from community hospitals, university hospitals and commercial laboratories. They found kappa coefficients of 0.888 in comparison with university hospitals, 0.465 with community hospitals and 0.419 with commercial labs and concluded there is a need for greater uniformity in the examination of biopsies. Niveloni *et al*[23] also recognised the discordance between academic histopathologists and more general histopathologists by observing that 12 of 59 cases diagnosed in community practises were determined to be misdiagnosed by an expert.

Corazza *et al*[20] compared the reports of six pathologists over 60 patients using both the Marsh–Oberhuber and Corazza–Villanacci grading schemes and found kappa coefficients of 0.35 and 0.55, respectively. It is worth noting that the Corazza–Villanacci grading system has fewer categories than the Marsh–Oberhuber scheme, making it more likely to yield better agreement.

Using 114 patients and five pathologists, Picarelli *et al*[24] reported kappa coefficients of 0.546 for agreement on villous–crypt ratios (<3 or ≥3), 0.406 for identifying intraepithelial lymphocytosis (based on classifications of above and below the threshold of 25 intraepithelial lymphocytes per 100 epithelial enterocytes), and 0.652 for classifications using the Marsh–Oberhuber scheme.

Eigner *et al*[25] examined 53 patients with CD diagnoses who were under suspicion of misdiagnosis. After an experienced pathologist reviewed biopsies from these cases, they found a kappa coefficient of 0.072, which corresponds to near-random agreement. The positive or negative CD status was determined using the Marsh–Oberhuber scheme. It is plausible that this near-random level of agreement is due to the fact that the cases were suspected to be misdiagnoses.

Inspired by a striking near 40-fold difference in incidence rates of CD between Denmark and Sweden,[26 27] Weile *et al*[28] investigated the interobserver agreement between Danish and Swedish Pathologists using 93 biopsies from 73 children. When comparing between three pathologists of 'moderate to substantial' experience, Weile *et al*[28] found kappa values of 0.57≤κ≤0.75, and in a comparison between the studies' pathologists and the existing diagnoses, found kappa values of 0.53≤κ≤0.57.[28] Weile *et al*[28] thus concluded there is no difference between the reporting of Swedish and Danish pathologists.

There are many other studies exploring concordance in the histological interpretation of duodenal biopsies in the context of CD, a systematic review of which is beyond the scope of this writing.[29–35] While there is considerable literature examining the interobserver agreement in CD diagnosis, there is a general lack of standardisation in the studies' designs.

Recently, Laohawetwanit *et al*[36] presented the findings of a global survey of pathologists' views on online digital pathology and in particular the use of digital whole-slide images (WSIs). They found that about two-thirds of all pathologists had no concern regarding the use of virtual slides for educational purposes and viewed them as an acceptable substitute for glass slides. Similarly, the Royal College of Pathologists states that 'Digital pathology is a technology which has the potential to transform the way pathologists work'.[37]

In this study, 13 pathologists classified 100 duodenal biopsies in the form of digitised WSIs as showing features of CD, indeterminate enteropathy or normal tissue, without any additional clinical information or blood results. The pathologists then examined the same (re-anonymised) cases in the presence of additional metadata data—namely IgA tTG and haemoglobin (Hb). To our knowledge, there are no other studies which investigate the general concordance in digital duodenal biopsy classifications; nor are there any digital or glass slide review duodenal biopsy concordance studies that use such a large number of pathologists. Finally, we believe we are

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

the first to analyse the effect of including additional data on the quality of the diagnosis.

METHODS AND MATERIALS

Data

One hundred H&E-stained duodenal (D2) biopsies were obtained from the Heart of England NHS Foundation Trust Hospital, Birmingham, UK and scanned on a Roche Ventana iScan HT at 40× objective magnification, which corresponds to a spatial resolution of 0.25 µm per pixel (note: the spatial resolution quoted at 40× magnification varies with scanner manufacturer).

The biopsies were classified as normal (n=40), CD/gluten sensitive enteropathy (n=40) or indeterminate enteropathy (n=20) based on a review of their histology, tTG/EMA serology and Hb level, and their clinical presentation. The participants were not made aware of the relative abundance of each category.

The WSIs were obtained by scanning a single H&E-stained level from cases with known diagnoses, made previously on a combined review of the patients' histology, serology and clinical presentation. In order to increase the total size of the dataset while keeping costs reasonable, we chose to have one well-chosen level per biopsy.

Instructions to participating pathologists

Thirteen specialist GI consultant pathologists, including four who had experience in digital reporting prior to this study, were informed the biopsies had been classified as normal, positive for CD/gluten sensitive enteropathy or indeterminate enteropathy (but not the number of instances of each class). The participants were instructed to interpret and diagnose each case in the same way they would in their own, standard, national health service reporting practice. The study was organised in two phases:

- *Phase 1*. The GI pathologists independently examined the 100 WSIs in the absence of any serological or clinical data.
- *Phase 2*. The same pathologists repeated the study (with re-anonymised images) with the inclusion of additional data (IgA tTG and Hb).

All cases had Hb available, but in 37 cases, the IgA tTG data were missing. Rather than carefully picking 100 biopsies with Hb and tTG data, we aimed to include a set of biopsies that closely resemble real-world data.

A lack of standardised reporting practices exists across medical centres for cases not classified as normal or CD, leading pathologists to employ diverse terminology in their routine analysis. We adopt the term 'indeterminate enteropathy' to capture the varied terms regularly used by pathologists including 'non-specific (chronic) inflammation', 'active inflammation', 'non-specific duodenitis', 'acute duodenitis' or 'partial villous atrophy'.

WSI access

The pathologists accessed the WSIs using the Comparative Pathology Workbench (CPW), developed at the University of Edinburgh.[38–40] The CPW is an integrated tool for spatial data annotation and analysis and allows easy comparison WSIs.

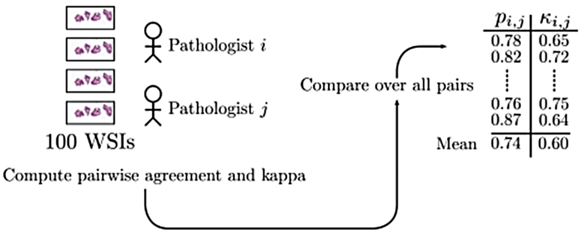

Figure 1 Overview of the interobserver agreement methodology: for each distinct pair of pathologists, i and j , we measure the interobserver agreement across the 100 WSIs, before averaging over all possible pairs. WSIs, whole-slide images.

Analysis

To measure the interobserver agreement, we collated the independent answers from each pathologist. For each distinct pair of observers, i and j, we compared their classifications across the 100 WSIs and computed the probability that they should agree $p_{i,j}$ and the corresponding Cohen's kappa coefficient $\kappa_{i,j}$.[41] For clarity, the agreement between two observers, i and j, measured using Cohen's kappa coefficient, is defined as

$$\kappa_{i,j} = 1 - \frac{1-p_{i,j}}{1-p_e}$$

where $p_{i,j}$ is the observed probability of the two observers agreeing and p_e is the theoretical probability that the observers should agree by virtue of chance.[41] We repeated this process for every distinct pair of observers, before obtaining estimates of the mean probability of agreement and the mean kappa coefficient by averaging over the 76 total possible pairs (figure 1).

RESULTS

Interobserver agreement statistics

We first considered the interobserver agreement in phase one, where the pathologists independently examined the WSIs in the absence of any serological or clinical data. The mean probability of two observers agreeing on a given diagnosis was 0.73 (±0.08) which corresponded to a mean Cohen's kappa coefficient of 0.59 (±0.11) (figure 2A).

We next considered the interobserver agreement in phase 2, where the same pathologists examined the (re-anonymised) WSIs with additional data—namely IgA tTG and Hb. However, with additional data, the probability of agreement increased to 0.80 (±0.06) and the Cohen's kappa to 0.67 (±0.09) (figure 2B). We tested the statistical significance of the observed increase in the means of the probability of agreement and Cohen's kappa and found p values of order 10^{-10} and 10^{-9}, respectively (online supplemental Appendix A table 1).

We further highlighted the varying interobserver agreement by disaggregating the statistics by observer and phase of the study in online supplemental Appendix

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

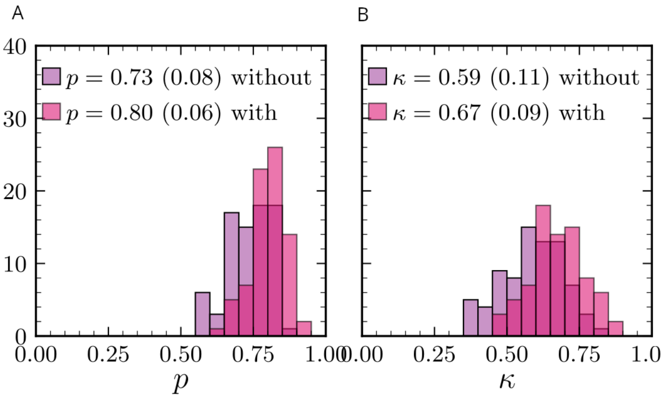

Figure 2 Probability distributions of the agreement probability (A) and Cohen's kappa (B) comparing the concordance of the diagnosis made with and without metadata data. We show that the agreement was higher when the pathologists had access to the additional IgA tTg and Hb data. Note: the legends show mean (SD). Hb, haemoglobin; IgA tTG, immunoglobulin A tissue transglutaminase.

B table 2. In short, the inclusion of IgA tTG and Hb data improved the interobserver agreement (while reducing the SD by about 10%) in the histological interpretation of duodenal biopsies.

Pathologists' use of categories

We also examined how frequently the observers opted for each option (normal, CD and indeterminate enteropathy) and, intriguingly, found a marked variation between individual pathologists.

Figure 3A shows the frequency with which each observer selected each option. Strikingly, observer 'm' determined 23/100 WSIs to be normal, while observer 'a' judged 63/100 as such. In the case of indeterminate enteropathy, observer 'a' identified 6/100, while observer 'm' reported 37/100 cases to be indeterminate enteropathy. Finally, observer 'l' determined 54/100 WSIs to be cases of CD, yet observer 'g' only 20/100. These strong contrasts clearly highlight the lack of uniformity in the assessment of the histological features of duodenal biopsies by GI pathologists who routinely report such specimens.

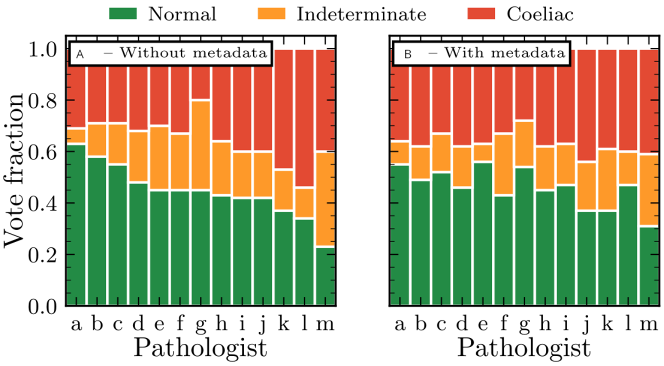

Figure 3 Summary of the voting patterns disaggregated by the phase of the study and individual pathologist. (A) Summarises the decisions made in the absence of serological data and (B) with the inclusion of the additional data. This clearly highlights the lack of uniformity in the assessment of the histological features of duodenal biopsies by specialist GI pathologists. GI, gastrointestinal.

In figure 3B, we illustrate the effect of additional data on the frequency of each diagnosis. We show that, without metadata the observer most likely to interpret a slide as normal did so in 40 cases more than the observer least likely to do so. However, with additional data (Hb and IgA tTG), the gap between the greatest and smallest number of normal votes narrowed to 24 cases. Similarly, in the case of indeterminate enteropathy, the range in the number of votes decreased from 24 to 21 cases, and in the case of CD, from 34 to 16. The inclusion of the additional data therefore significantly decreased the range in the number of votes for each category. This result is in line with the finding that the interobserver agreement increases with the inclusion of supporting metadata.

We further showed that most cases of disagreement are between indeterminate diagnosis and either normal or CD. In phase 1 of the study, 12% of all cases include one pathologist diagnosing a WSI as indeterminate and the other as CD. Similarly in 10% of all cases, we got a normal and an indeterminate classification. In contrast, in only 4% of all cases, one pathologist diagnosed a WSI as normal and the other as CD. With the inclusion of additional metadata, the number of normal-CD disagreements reduced even further to 2%. We illustrated the full confusion matrices for both phases of the study in online supplemental Appendix D table 4.

Per case agreement

As shown in figure 4, a significant number of cases have a 100% agreement between all 13 pathologists. In the absence of serological data, 18 cases were diagnosed as normal by all 13 pathologists, 15 as CD and 1 as indeterminate. In phase 2 of the study, when the pathologists had access to Hb and tTG data, the number of normal and CD cases with 100% agreement increased to 22 and 20, respectively. We analysed mean agreement on cases with missing tTG in Appendix F and compare it against cases with full serology, observing that the inclusion of Hb and tTG results in a higher increase in concordance when contrasted with the addition of Hb alone.

Metadata-dependent intraobserver agreement

Finally, we compared the pathologists' classifications from each phase and thus measured their 'self-agreement' between making diagnosis with and without metadata (online supplemental Appendix C table 3 and online supplemental Appendix D table 5). The mean probability that an observer's determination for a given WSI remained unchanged with and without additional data is 0.79 (±0.05), with a corresponding Cohen's kappa of 0.66 (±0.08). It is therefore clear that the additional IgA tTG and Hb data play a significant role in individuals' interpretation of WSIs.

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

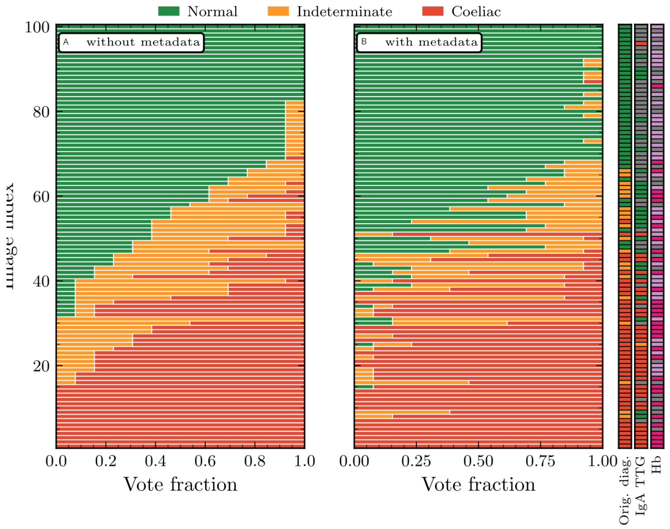

Figure 4 Illustrating the diagnosis made by each pathologist disaggregated by case for both phases of the study. (A) In phase 1 of the study without any available metadata. 18 cases were diagnosed as normal by all 13 pathologists, 15 as CD and 1 as indeterminate. (B) In phase 2 of the study with available Hb and tTG data 22 cases were diagnosed as normal by all 13 pathologists, 20 as CD and 0 as indeterminate. Each row represents a single case. CD, coeliac disease; Hb, haemoglobin; IgA tTG, immunoglobulin A tissue transglutaminase.

Pathologists' prior digital experience

Four pathologists routinely reported digitally in their national health practice at the time of the study. For the pathologists with prior digital reporting experience, we observe mean agreement and kappa coefficient of 0.59 and 0.74 in phase 1 of the study, which increased to 0.70 and 0.82, respectively, in the presence of metadata. For the other group of pathologists, we observe a mean agreement/kappa coefficient of 0.58 and 0.73 without metadata, which increased to 0.66 and 0.79, respectively, in phase 2. We thus observe no meaningful difference between the pathologists with and without prior digital reporting experience.

DISCUSSIONS AND CONCLUSIONS

Summary

We investigated the interobserver agreement over the 78 two-observer permutations in a group of 13 GI pathologists, each of whom classified 100 WSIs of H&E-stained duodenal biopsies without any serological, clinical or genetic context in a purely digital setting. We included cases previously classified as normal (n=40), CD/gluten-sensitive enteropathy (n=40) or indeterminate (n=20). The mean probability of two observers agreeing on a given diagnosis was 0.73 (±0.08) with a corresponding Cohen's kappa coefficient of 0.59 (±0.11).

Next, we evaluated the importance of IgA tTG and Hb data in coeliac diagnosis by having the pathologists examine the (re-anonymised) WSIs with these additional data. The added data increased the probability of two observers agreeing by about 10% to 0.80 (±0.06). The corresponding kappa coefficient also increased by over 10% to 0.067 (±0.09). In the case of both the probability of agreement and Cohen's kappa, the increase in agreement after the inclusion of the additional data was statistically significant (p~10^{-10} and p~10^{-9}, respectively).

Potential for bias

Despite the increasing uptake in digital pathology, there are varying levels of experience in reporting WSIs; many pathologists still routinely use optical microscopes. Moreover, as the majority of duodenal biopsies in routine clinical practice are diagnosed as normal, we enriched our dataset with cases of CD and cases reported to show evidence of indeterminate enteropathy. The relative abundance of each class likely has a significant impact on the probability of two observers agreeing, making it imperative to consider Cohen's kappa coefficient, which is a more robust metric for agreement.

Another important caveat to consider is that in routine practice, pathologists can request additional levels to be cut from the biopsy if they feel a specimen is of insufficient quality or unlikely to be fully representative of the material, whereas in this study the participants were restricted to only a single level per case.

Furthermore, the IgA tTG and Hb data were incomplete, so when the pathologists re-examined the WSIs some cases were missing data. Even with this minor compromise, the observed increase in interobserver agreement was statistically significant, so it is highly unlikely the small amount of missing data would qualitatively affect our findings.

Future work

This work raises a number of important questions. First, it would be interesting to investigate whether the level of agreement differs if the observers instead examined slides using optical microscopes: it is well known that the digitisation process is imperfect and can give rise to regions of blur and other artefacts which hinder the inspection of a slide. Unfortunately, this was outside the scope of our study, due to logistical challenges in sending the same slides to more than ten hospitals in different countries.

Second, in this study the observers examined each case independently, however, if they were to confer on each case in small groups, before deciding on a diagnosis by majority vote, it would be interesting to know if the level of agreement changes.

Third, while serological tests seem, on the surface, more objective in comparison to the histological interpretation of biopsies, it is important to consider that studies which examine the diagnostic utility of serological tests for CD validate such tests against histology, meaning they are necessarily biased. Even though serological data increase the interobserver agreement, they do not necessarily improve the accuracy of diagnosis.

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

Fourth, it would have been interesting to perform a true intraobserver agreement study where the pathologists observe the same 100 biopsies in identical settings (without serological data) to see what variation exists in diagnosis between separate examinations by the same pathologists. However, it was outside the scope of this study due to the practical challenge of getting the very busy pathologists (in a country with a significant shortage of pathologists) to look at the 100 biopsies for a third time.

Conclusion

There is a clear and unmet need to address the non-uniform standards that GI pathologists apply in the diagnosis of CD, for example, by developing a more objective test for CD (such as algorithmic approaches to image analysis[42–47]). Some have also argued that diagnosis becomes more reproducible with the incorporation of manual software tools into reporting processes,[48] but such approaches are far from standard practice.

The era of digital pathology brings opportunities automating disease diagnosis and the creation of decision support tools to aid pathologists in routine practice. However, the challenge of the low diagnostic concordance between pathologists, when examining duodenal biopsies, highlights the need to develop very carefully curated datasets, in which diagnostic accuracy (ground truth) is optimised, ideally by considering histopathology, serological results, haemoglobin level and clinical data together.

Author affiliations
[1]Department of Pathology, University of Cambridge, Cambridge, UK
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
[3]Lyzeum Ltd, Cambridge, UK
[4]Department of Pathology, The University of Edinburgh College of Medicine and Veterinary Medicine, Edinburgh, UK
[5]Division of Medical Education, The University of Manchester, Manchester, UK
[6]Department of Histopathology, Manchester University NHS Foundation Trust, Manchester, UK
[7]Department of Diagnostic and Molecular Pathology, Royal Cornwall Hospitals NHS Trust, Truro, UK
[8]University Hospitals Plymouth NHS Trust, Plymouth, UK
[9]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[10]Department of Pathology, Akershus University Hospital, Nordbyhagen, Norway
[11]Institute of Clinical Medicine, University of Oslo, Nordbyhagen, Norway
[12]Department of Cellular Pathology, Oxford University Hospitals NHS foundation Trust, Oxford, UK
[13]Department of Histopathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK
[14]Department of Pathology, Western General Hospital, Edinburgh, UK
[15]Department of Cellular Pathology, Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[16]MRC Institute of Metabolic Science, Wellcome Trust, Cambridge, UK
[17]Department of Pathology, Queen Elizabeth University Hospital, Glasgow, UK
[18]University Hospital of North Tees, North Tees and Hartlepool NHS Foundation Trust, Stockton on Tees, UK
[19]Division of Pathology, University of Edinburgh, Edinburgh, UK

Acknowledgements JD, MJA and ES acknowledge Graham Snudden for organisational support. JD, BS, FJ, MJA and ES are grateful to the 17 GI pathology experts for agreeing to participate, and to MW for facilitating virtual access to the WSIs, all of whom made this work possible.

Contributors JD, FJ and BS conducted and tested all of the analysis. JD and ES drafted the original revision of this manuscript, which was heavily contributed to by FJ, BS and MJA. MW set up and facilitated access virtual access to the WSIs for each participant. All authors, with the exception of JD, BS, MW, FJ, MJA and ES, reported on the 100 duodenal biopsies. ES and MJA conceptualised this study and guided the analysis. BS, FJ, MJA and ES reviewed and proposed revisions to this manuscript, before it was circulated with all co-authors, who then had the opportunity to do the same. JD is the guarantor for this study.

Funding JD and ES acknowledge Coeliac UK and Innovate UK (grant INOV03-19 to ES). FJ acknowledges financial support from the Cambridge Centre for Data-Driven Discovery (C2D3). ES acknowledges a Pathological Society Consultant's PumpPriming grant—01 April 2016; Grant Reference No: 1084 'Developing digital analytical algorithms for coeliac disease diagnosis: a paradigm for epithelial/inflammatory pathology'. BS acknowledges a Pathological Society PhD studentship. The Comparative Pathology Workbench has been further developed for use by the Gut Cell Atlas Crohn's Disease Consortium funded by The Leona M. and Harry B. Helmsley Charitable Trust and is supported by a grant from Helmsley to the University of Edinburgh.

Competing interests ES and MS are shareholders in Lyzeum Ltd, by whom JD was employed at the time of this work.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved. All slide scans (and accompanying fully anonymised patient data) were obtained with full ethical approval (IRAS: 162057; PI: Dr E Soilleux). Organisation: South Central – Oxford A (formerly known as Oxfordshire Research Ethics Committee A). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The raw data, along with the code and instructions for reproducing all of the analysis and figures presented in this work are available in THIS GITLAB REPOSITORY (https://gitlab.developers.cam.ac.uk/path/soilleux/soilleux-group/cd-inter-observer-agreement). We are not at liberty to share the WSIs, however.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Florian Jaeckle http://orcid.org/0009-0002-4607-6476
Betania Mahler-Araujo http://orcid.org/0000-0001-7952-0292

REFERENCES

1 Adams F. *The extant works of Aretaeus, the Cappadocian*. London: London Sydenham society, 1856. Available: https://archive.org/details/b21510271
2 Paveley WF. From Aretaeus to Crosby: A history of Coeliac disease. *BMJ* 1988;297:1646–9.
3 Green PHR, Cellier C. Celiac disease. *N Engl J Med* 2007;357:1731–43.
4 Lebwohl B, Sanders DS, Green PHR. Coeliac disease. *The Lancet* 2018;391:70–81.
5 Losowsky MS. A history of Coeliac disease. *Dig Dis* 2008;26:112–20.
6 Corazza GR, Villanacci V. Coeliac disease. *J Clin Pathol* 2005;58:573–4.

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.

7 Caio G, Volta U, Sapone A, *et al*. Celiac disease: A comprehensive current review. *BMC Med* 2019;17:142.
8 Singh P, Arora A, Strand TA, *et al*. Global prevalence of celiac disease: systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology* 2018;16:823–836.
9 O'Reilly D, Murphy J, McLaughlin J, *et al*. The prevalence of Coeliac disease and cystic fibrosis in Ireland, Scotland, and England and Wales. *Int J Epidemiol* 1974;3:247–51.
10 Johnston S, Watson R, McMillan S, *et al*. Prevalence of Coeliac disease in Northern Ireland. *The Lancet* 1997;350:1370.
11 Dydensborg S, Toftedal P, Biaggi M, *et al*. Increasing prevalence of Coeliac disease in Denmark: A linkage study combining national registries. *Acta Paediatr* 2012;101:179–84.
12 White LE, Merrick VM, Bannerman E, *et al*. The rising incidence of celiac disease in Scotland. *Pediatrics* 2013;132:e924–31.
13 Catassi C, Fasano A. Celiac disease diagnosis: simple rules are better than complicated Algorithms. *Am J Med* 2010;123:691–3.
14 Husby S, Koletzko S, Korponay-Szabó IR, *et al*. European society for pediatric Gastroenterology, Hepatology, and nutrition guidelines for the diagnosis of Coeliac disease. *J Pediatr Gastroenterol Nutr* 2012;54:136–60.
15 guidelines N. NICE guidelines | NICE guidance | Our programmes | What we do | About | NICE, Available: 2018.https://www.nice.org.uk/About/What-we-do/Our-Programmes/NICE-guidance/NICE-guidelines [Accessed 17 May 2022].
16 National Insitute for Health Care and Excellence. Recommendations | coeliac disease: recognition, assessment and management | guidance. Available: 2015.https://www.nice.org.uk/guidance/ng20/chapter/Recommendations{\#}recognition-of-coeliac-disease https://www.nice.org.uk/guidance/ng20/chapter/Recommendations{\#}serological-testing-for-coeliac-disease [Accessed 17 May 2022].
17 Ludvigsson JF, Bai JC, Biagi F, *et al*. Diagnosis and management of adult coeliac disease: guidelines from the British society of Gastroenterology. *Gut* 2014;63:1210–28.
18 Marsh MN. Gluten, major Histocompatibility complex, and the small intestine. A molecular and Immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue. *Gastroenterology* 1992;102:330–54.
19 Georg O, Gerhard G, Harald V. The Histopathology of Coeliac disease: time for a standardized report scheme for Pathologists. *Eur J Gastroenterol Hepatol* 1999;11:1185.
20 Corazza GR, Villanacci V, Zambelli C, *et al*. Comparison of the Interobserver reproducibility with different histologic criteria used in celiac disease. *Clin Gastroenterol Hepatol* 2007;5:838–43.
21 Ensari A. Gluten-sensitive Enteropathy (celiac disease): controversies in diagnosis and classification. *Archives of Pathology & Laboratory Medicine* 2010;134:826–36.
22 Arguelles-Grande C, Tennyson CA, Lewis SK, *et al*. Variability in small bowel Histopathology reporting between different pathology practice settings: impact on the diagnosis of Coeliac disease. *J Clin Pathol* 2012;65:242–7.
23 Niveloni SI, Cabanne AM, Vázquez H, *et al*. Experts' assess the accuracy of celiac disease diagnosis performed in the community setting. *Gastroenterology* 2012;142:S–183.
24 Picarelli A, Borghini R, Donato G, *et al*. Weaknesses of histological analysis in celiac disease diagnosis: new possible scenarios. *Scand J Gastroenterol* 2014;49:1318–24.
25 Eigner W, Wrba F, Chott A, *et al*. Early recognition of possible pitfalls in histological diagnosis of celiac disease. *Scand J Gastroenterol* 2015;50:1088–93.
26 Bodé S, Gudmand-Høyer E. Incidence and prevalence of adult coeliac disease within a defined geographic area in Denmark. *Scand J Gastroenterol* 1996;31:694–9.
27 Sjöberg K, Eriksson S. Regional differences in coeliac disease prevalence in Scandinavia *Scand J Gastroenterol* 1999;34:41–5.
28 Weile B, Hansen BF, Hägerstrand I, *et al*. Interobserver variation in diagnosing coeliac disease. A joint study by Danish and Swedish Pathologists. *APMIS* 2000;108:380–4.
29 Mubarak A, Nikkels P, Houwen R, *et al*. Reproducibility of the histological diagnosis of celiac disease. *Scand J Gastroenterol* 2011;46:1065–73.
30 Mahatma Gandhi memorial medical College, Indore (M.P) India, Ghanghoria S, Sharma S, *et al*. Celiac disease: comparison of Oberhuber classification and Corazza- Villanacci classification. *APALM* 2019;6:A135–140.
31 Kaur Bilkhoo H, Ducruet T, Marchand V, *et al*. Revisiting pathological criteria for earlier diagnosis of Coeliac disease. *J Pediatr Gastroenterol Nutr* 2016;62:734–8.
32 van Wanrooij RLJ, Müller DMJ, Neefjes-Borst EA, *et al*. Optimal strategies to identify aberrant intra-epithelial lymphocytes in refractory coeliac disease. *J Clin Immunol* 2014;34:828–35.
33 Montén C, Bjelkenkrantz K, Gudjonsdottir AH, *et al*. Validity of Histology for the diagnosis of Paediatric coeliac disease: A Swedish Multicentre study. *Scand J Gastroenterol* 2016;51:427–33.
34 Webb C, Halvarsson B, Norström F, *et al*. Accuracy in celiac disease diagnostics by controlling the small-bowel biopsy process. *J Pediatr Gastroenterol Nutr* 2011;52:549–53.
35 Willington R, Lashmar V, Benes K, *et al*. PTH-184 push Enteroscopy leads to a change in diagnosis in the majority of patients with positive coeliac Serology and negative Duodenal biopsy. *Gut* 2013;62(Suppl 1):A286.
36 Laohawetwanit T, Gonzalez RS, Bychkov A. Learning at a distance: results of an international survey on the adoption of virtual conferences and whole slide imaging by Pathologists. *J Clin Pathol* 2023:jcp-2023-208912.
37 The Royal college of Pathologists. *Position statement from the Royal College of Pathologists (RCPath) on Digital Pathology and Artificial Intelligence (AI)*. Available: https://www.rcpath.org/static/90e5e248-4ad3-4d61-8247223f9faffc80/RCPath-AI-position-statement-2022.pdf
38 Wicks MN, Glinka M, Hill B, *et al*. *Comparative Pathology Workbench*. The University of Edinburgh, 2021.
39 Wicks MN, Glinka M, Hill B, *et al*. The comparative pathology workbench: interactive visual Analytics for BIOMEDICAL data. *J Pathol Inform* 2023;14:100328.
40 Wicks MN, Glinka M, Hill B, *et al*. Comparative pathology workbench. *GitHub Repository* 2023.
41 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37–46.
42 Denholm J, Schreiber BA, Evans SC, *et al*. Multiple-instance-learning-based detection of Coeliac disease in histological whole-slide images. *J Pathol Inform* 2022;13:100151.
43 Kowsari K, Sali R, Khan MN, *et al*. Diagnosis of celiac disease and environmental enteropathy on biopsy images using color balancing on convolutional neural networks. *Proc Futur Technol Conf FTC (2019)* 2020;1069:750–65.
44 Wei JW, Wei JW, Jackson CR, *et al*. Automated detection of celiac disease on Duodenal biopsy slides: A deep learning approach. *J Pathol Inform* 2019;10:7.
45 Sali R, Ehsan L, Kowsari K, *et al*. Celiacnet: celiac disease severity diagnosis on Duodenal histopathological images using deep residual networks. Proceedings (IEEE Int Conf Bioinformatics Biomed 2019;2019:962–7.
46 Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;25:1301–9.
47 Schreiber BA, Denholm J, Gilbey JD, *et al*. Stain normalization gives greater Generalizability than stain Jittering in neural network training for the classification of Coeliac disease in Duodenal biopsy whole slide images. *J Pathol Inform* 2023;14:100324.
48 Das P, Gahlot GP, Singh A, *et al*. Quantitative histology-based classification system for assessment of the intestinal Mucosal histological changes in patients with celiac disease. *Intest Res* 2019;17:387–97.

BMJ Open Gastroenterol: first published as 10.1136/bmjgast-2023-001252 on 1 February 2024. Downloaded from http://bmjopengastro.bmj.com/ on February 3, 2024 by guest. Protected by copyright.