Phenome of coeliac disease vs. inflammatory bowel disease

Moritz Kleinjans1,3, Carolin V. Schneider1,2,3, Tony Bruns1 & Pavel Strnad1,*

Coeliac disease (CeD) is characterized by gliadin-induced intestinal inflammation appearing in genetically susceptible individuals, such as HLA-DQ2.5 carriers. CeD, as well as other chronic intestinal disorders, such as Crohn’s disease (CD) and ulcerative colitis, has been associated with increased morbidity and mortality, but the causes are unknown. We systematically analysed CeD-associated diagnoses and compared them to conditions enriched in subjects with CD/UC as well as in HLA-DQ2.5 carriers without CeD. We compared the overall and cause-specific mortality and morbidity of 3,001 patients with CeD, 2,020 with CD, 4,399 with UC and 492,200 controls in the community-based UK Biobank. Disease-specific phenotypes were assessed with the multivariable Phenome Wide Association Study (PheWAS) method. Associations were adjusted for age, sex and body mass index. All disease groups displayed higher overall mortality than controls (CD: aHR = 1.91 [1.70–2.17]; UC: aHR = 1.32 [1.20–1.46]; CeD: aHR = 1.38 [1.22–1.55]). Cardiovascular and cancer-related deaths were responsible for the majority of fatalities. PheWAS analysis revealed 166 Phecodes overrepresented in all three disorders, whereas only ~ 20% of enriched Phecodes were disease specific. Seven of the 58 identified CeD-specific Phecodes were enriched in individuals homozygous for HLA-DQ2.5 without diagnosed CeD. Four out of these seven Phecodes and eight out of 19 HLA-DQ2.5 specific Phecodes were more common in homozygous HLA-DQ2.5 subjects with vs. without CeD, highlighting the interplay between genetics and diagnosis-related factors. Our study illustrates that the morbidity and mortality in CeD share similarities with CD/UC, while the CeD-restricted conditions might be driven by both inherited and acquired factors.

Chronic inflammatory intestinal diseases constitute a great burden for patients and health systems1,2. The most relevant entities include Crohn’s disease (CD), ulcerative colitis (UC) and coeliac disease (CeD). CD and UC share many similarities and are therefore often summarized as inflammatory bowel disease (IBD). They have a multifactorial aetiology characterized by a dysfunctional intestinal barrier and dysregulated immune responses3–5. While UC is limited to the colon, CD can affect the entire gastrointestinal tract and typically includes the terminal ileum4,5. Patients with CD often develop complications such as abscesses and fistulas, while colorectal carcinoma constitutes a characteristic complication of UC4.

CeD is an immune-mediated enteropathy causing chronic duodenal inflammation in genetically predisposed individuals. It is triggered by exposure to gliadin, a widely spread dietary component2. In addition to intestinal symptoms such as abdominal pain, diarrhoea, and malabsorption, CeD is also associated with extraintestinal comorbidities, including cardiovascular and haematologic diseases, as well as many diseases from the autoimmune spectrum6–8. The main genetic predisposition for the development of CeD is the presence of susceptible human leukocyte antigens (HLAs), i.e., cell-surface proteins responsible for the regulation of the immune system. Among them, HLA-DQ2 or -DQ8 are able to bind gluten-derived peptides and present them to CD4+ T cells9. Most patients with CeD carry one or more of the disease-related HLA alleles DQ2.5, DQ2.2, or DQ8, but the presence of these variants is not sufficient to develop CeD10,11. The DQ2.5 variant is the most important genetic factor, occurring in approximately 90% of CeD patients9. Since an apparent gene-dose effect exists, patients carrying the homozygous HLA-DQ2.5 allele are at particularly high risk for CeD10,11. However, other factors, including non-HLA genetic risk factors and changes in the microbiome, might play a role in CeD, and the same is true for CD/UC10–12.

Although CeD is deemed less life-threatening than CD and UC, several studies suggest that it is associated with excess mortality13–16. The underlying mechanisms remain unclear, but they might be related to the presence of acquired (i.e., diagnosis-related) factors, genetic background, or both. To address this issue, we used the UK...
Biobank and compared the overall and cause-specific mortality and morbidity of patients with CeD and IBD. The goal of our study was to identify how many CeD-associated disorders are CeD-specific and how many are shared with CD/UC. Finally, we wondered whether the CeD-specific Phecodes are related to the genetic background of CeD subjects or are instead associated with diagnosis-related factors.

### Results

#### Overall- and cause-specific mortality of analysed intestinal diseases.

Among the 502,488 individuals recruited to UKB, 3001 were diagnosed with coeliac CeD, 4,399 with UC, and 2020 with CD, while 868 individuals had none of these diagnoses and were used as a reference group (Supplementary Fig. S2). All analyses were adjusted for age, sex and body mass index. For details, see Supplementary Table S4. 

Table 1. Baseline characteristics of the study population. The reference group consists of individuals without CeD’s disease, ulcerative colitis, or coeliac disease. BMI body mass index.

| Cause of death       | Coeliac disease [K90] (n = 3001) | Ulcerative colitis [K51] (n = 4399) | Crohn’s disease [K50] (n = 2020) |
|----------------------|----------------------------------|-----------------------------------|----------------------------------|
|                      | Reference (n = 492,200)          |                                   |                                  |
| Age (years)          | 56.5 ± 8.1                       | 57.8 ± 7.8                        | 57.6 ± 7.9                       | 56.8 ± 8.1                     |
| Males                | 224,490 (45.6%)                  | 1046 (34.9%)                      | 2293 (52.1%)                     | 872 (43.2%)                    |
| BMI (kg/m²)          | 27.4 ± 4.8                       | 26.2 ± 4.8                        | 27.6 ± 4.7                       | 27.2 ± 5.0                     |

Table 2. Overall and disease-specific mortality in subjects with analysed intestinal disorders. Multivariable hazard ratios were determined for patients with coeliac disease, ulcerative colitis, and Crohn’s disease compared to the reference group that displays none of these conditions. All analyses were adjusted for age, sex and body mass index. For details, see Supplementary Table S4. PYs person-years, CI confidence interval, HR hazard ratio. Significant values are in bold.

| Cause of death       | Coeliac disease [K90] (n = 3001) | Ulcerative colitis [K51] (n = 4399) | Crohn’s disease [K50] (n = 2020) |
|----------------------|----------------------------------|-----------------------------------|----------------------------------|
|                      | n deaths/1000 PYs | HR 95%-CI | n deaths/1000 PYs | HR 95%-CI | n deaths/1000 PYs | HR 95%-CI |
| Overall mortality    | 271 7.72        | 1.38 1.22–1.55 | 437 8.49         | 1.32 1.20–1.46 | 250 10.67        | 1.91 1.70–2.17 |
| Malignancies (C00-C97)| 129 3.67       | 1.22 1.03–1.45 | 205 3.98         | 1.25 1.09–1.43 | 108 4.61         | 1.60 1.33–1.93 |
| Digestive diseases (K00-K93) | 17 0.48   | 2.58 1.60–4.16 | 37 0.72          | 3.14 2.26–4.35 | 30 1.28          | 6.45 4.49–9.26 |
| Cardiovascular diseases (I00-I99) | 51 1.45 | 1.41 1.07–1.86 | 88 1.71          | 1.21 0.98–1.50 | 50 2.13          | 1.90 1.44–2.52 |
| Respiratory diseases (J00-J99) | 25 0.71   | 1.76 1.18–2.61 | 32 0.62          | 1.32 0.93–1.88 | 21 0.90          | 2.29 1.49–3.52 |

PheWAS analysis reveals disease-specific phenotypes. To decipher whether the observed comorbidities represent a conserved response to chronic intestinal injury or whether they constitute unique, disease-specific traits, we performed multivariable PheWAS analysis comparing the different intestinal diseases with the reference group (Fig. 1). Among all investigated intestinal disorders, an enrichment of Phecodes for gastrointestinal, haematopoietic and metabolic diseases, symptoms, and complications was observed. Osteoporosis, anaemia, and dermatological Phecodes were more prominent in patients with CeD than in patients with IBD (Fig. 1). The phenotypic overlap between the intestinal disorders became apparent in the Venn diagram, which revealed 166 Phecodes shared between all three disorders (Fig. 2). The number of shared Phecodes between CD and UC was significantly greater than that between CeD and one of the two inflammatory bowel disease subgroups (CeD and CD: 28; CeD and UC: 32; CD and UC: 68). The number of Phecodes specific to each disease was comparable (CeD: 58, UC: 59), while CD had the highest number of unique Phecodes (CD: 79) (Fig. 2).

Notably, only 18–23% of associated Phecodes were disease specific, while the others were shared by at least two disorders and 49–58% by all three disorders (Fig. 2).
Autoimmune phecodes are enriched in CeD. To delineate conditions associated with CeD, we assessed CeD-specific Phecodes (Fig. 3). Among them, Phecodes of well-established CeD-associated dermatologic manifestations such as “dermatitis herpetiformis” (OR = 122.2) and “bullous dermatoses” (OR = 23.6) displayed the highest odds ratios. Notably, Phecodes of autoimmune diseases were particularly prominent. They included diabetes and thyroid disorders (Fig. 3) and displayed odds ratios ranging from 3.6 (“other disorders of thyroid”) to 14.3 (“other immunological findings”). Several cardiovascular and ocular disorders were also enriched in patients with CeD, with odds ratios of approximately 2. The only Phecodes coding for malignancy that were

Figure 1. Conditions associated with CeD and IBD. Manhattan plots display Phecodes that are significantly over/underrepresented in individuals with coeliac disease (A), ulcerative colitis (B), or Crohn’s disease (C) compared with the reference group without these diagnoses. The ten most significant associations are shown. Upwards/downwards pointing triangles refer to Phecodes that are over/underrepresented. The black line indicates the significance level after Bonferroni adjustment for multiple testing. All analyses were adjusted for sex, age, and body mass index, and p values are displayed in a − log10 format. NOS not otherwise specified.
significantly enriched in patients with CeD were "pancreatic cancer" (OR = 3.1), "non-Hodgkin’s lymphoma" (OR = 2.0), and "malignancies of other lymphatic tissue" (OR = 1.9). As potential reasons for the relatively high respiratory disease-specific mortality, CeD individuals were more likely to harbour “other alveolar and parietoalveolar pneumonopathy” as well as “empyema and pneumothorax”. Finally, gastrointestinal diseases played only a minor role among the CeD-specific Phecodes. In contrast, digestive and genitourinary conditions were more commonly seen among CD-/UC-specific Phecodes, while infectious, autoimmune and ocular codes were relatively rare (Supplementary Figs. S3, S4). Moreover, digestive disorders were highly enriched among conditions that are overrepresented both in UC and CD but not CeD individuals (18 out of 68 Phecodes). Notably, the latter subgroup also contained 9 infection-related Phecodes (Supplementary Table S3).

Since Phecodes of autoimmune diseases constituted the most frequent disease spectrum enriched in subjects with CeD, we analysed whether autoimmune Phecodes were significantly associated with UC and CD as well. The corresponding Venn diagram revealed that five Phecodes were overrepresented in all intestinal disorders, with all of them being related to rheumatoid arthritis and psoriasis. One and two unique autoimmune Phecode(s) were more enriched in CD and UC individuals, respectively, compared to 13 in CeD subjects (Fig. 4).

**HLA-DQ2.5-dependent Phecodes and the impact of CeD diagnosis on the appearance of HLA-specific Phecodes.** As CeD is highly associated with specific HLA genotypes, we studied whether the underlying genetic background may contribute to the observed CeD-specific Phecodes. To this end, we focused on subjects with two DQ2.5 alleles, which are strongly predisposed to CeD (Supplementary Fig. S1B). A multivariable analysis restricted to patients without the diagnosis of CeD (Fig. 5A, Supplementary Fig. S1B) revealed that 19 Phecodes were overrepresented in individuals with two vs. 0–1 HLA-DQ2.5 alleles (Supplementary Table S1). The largest changes were observed in the categories “neoplasms” and “endocrine metabolic” (Fig. 5A). Significantly enriched Phecodes of malignant diseases comprised tumours of the respiratory system (OR = 1.6), “cancer of tongue” (OR = 2.6), “non-Hodgkin’s lymphoma” (OR = 1.9), “large cell lymphoma” (OR = 2.0), and “cancer of other lymphatic, histiocytic tissue” (OR = 1.7). Among Phecodes encoding endocrine diseases, diabetes mellitus type 1 and its complications as well as thyroid disorders, including “hypothyroidism”, “Graves’ disease”, and “thyrotoxicosis”, were the most prominent (Supplementary Table S1). In addition, “chronic hepatitis” (OR = 3.8) and “nonproliferative glomerulonephritis” (OR = 3.0) were also significantly enriched in subjects with a genetic predisposition to, but without the concomitant diagnosis of, CeD (Fig. 5A, Supplementary Table S1). Out of the 19 Phecodes enriched in noncoeliac DQ2.5 homozygotes, seven were also uniquely enriched in CeD, i.e., were overrepresented in CeD subjects but not UC/CD individuals (Fig. 5B). These included diabetes and its complications, “Graves’ disease”, “non-Hodgkin lymphoma,” and “cancer of lymphoid, histiocytic tissue”. This finding prompted us to investigate whether the observed HLA-associated Phecodes were further affected by the presence of CeD. To that end, we compared HLA-DQ2.5 homozygous individuals with and without the diagnosis of CeD (Fig. 6, Supplementary Table S2). Eight Phecodes were significantly enriched in the former group, while none of them was less frequent. These included “type 1 diabetes” (OR=3.1) together with...
its complications, “hypothyroidism” (OR = 2.4), “chronic hepatitis” (OR = 3.1), and “non-Hodgkin’s lymphoma” (OR = 2.8). In contrast, the frequency of respiratory tract cancer and thyrotoxicosis/Graves’ disease was similar in HLA-DQ2.5 homozygous subjects with and without the diagnosis of CeD (Fig. 6, Supplementary Table S2).

In conclusion, while some of the CeD-specific Phecodes were primarily driven by the HLA dosage, the majority might be HLA-independent or driven by a combination of genetic susceptibility and the presence of diagnosis-related factors.

**Discussion**

Our study demonstrated that all three analysed intestinal disorders are associated with overall excess mortality, with CD conferring the highest hazard ratio. Although this was suggested previously, the risks seen in the UKB cohort somewhat exceed the numbers reported in large meta-analyses. In all three diseases, cancer- and cardiovascular-related deaths accounted for the majority of the cases, and the specific death risks closely resembled the overall mortality for the corresponding intestinal disorder. While increased cancer-related mortality has been established for CD, the data on other disorders and the association between intestinal inflammation and cardiovascular mortality are conflicting, and further studies are needed to clarify the discrepant observations.

The excess digestive disease-related deaths that were particularly overrepresented in CD are well in line with published findings. They, however, accounted for only a minority of cases and did not explain the elevated overall death rate. In summary, our data suggest that the presence of chronic intestinal disease, irrespective of its...
Intestinal disease may also increase the susceptibility to other inflammatory diseases, such as rheumatoid arthritis and psoriasis, as these diseases are associated with all three intestinal disorders. Notably, IBD, psoriasis, and rheumatoid arthritis display alterations in similar inflammatory pathways involving, among others, Th17 cells\(^{35,36}\), and are targeted by comparable anti-inflammatory treatment strategies\(^{4,5,37,38}\). Although immunosuppressive drugs are not used for coeliac disease, the involvement of Th17 cells has been demonstrated\(^{39,40}\), and multiple reports describe an association between CeD and rheumatoid arthritis as well as psoriasis\(^{41,42}\). However, the observed association of all analysed intestinal disorders with rheumatoid arthritis might be confounded by the difficult discrimination from enteropathic arthritis\(^{43}\).

A major limitation of the study is its cohort design, which precludes the identification of causal relationships. Moreover, in the UK Biobank cohort, the diagnosis of CeD is based on ICD10 codes, and no histological data are available. The prevalence of CeD in the UK biobank is only approximately half of what would be expected from prevalence studies (~ 1%)\(^{44}\). This suggests that some cases remain undiagnosed and/or unreported, which is in line with previous population-based studies\(^{45}\). This can lead to the overrepresentation of more severe CeD cases and might be partly responsible for the high overall mortalities seen in our study. However, several well-performed studies used the same approach to define participants with CeD and observed a similar performance as other cohort studies\(^{46,47}\). In contrast, the prevalence of CD/UC subjects meets/exceeds the rates reported in other studies, which suggests a satisfactory diagnostic rate. Another important limitation is that due to the complexity of CeD, we were not able to address all potential contributors, such as the role of further genetic
variants or acquired factors, such as intestinal leakage, chronic immune cell activation, dietary composition or the microbiome. Because of the former, the importance of genetic factors might be larger than suggested by our analysis. Further studies are needed to delineate the exact origin of CeD-associated Phenotypes and to determine which conditions directly contribute to the increased mortality seen in CeD.

Methods

Study cohort. The 'UK Biobank' is a population-based cohort study that enrolled approximately 500,000 participants aged 37–73 years at 22 recruiting centres all over the United Kingdom. At the baseline visit (2006–2010), all participants gave consent for genotyping and data linkage to medical reports. Physical measurements and demographic information were obtained, and medical reports from 1996 onwards were used to record ICD-10 diagnoses. This source was used to identify patients with CeD (K90.0), UC (K51), and CD (K50) for our analysis.

The UK Biobank is linked to the national death register, which provided age at death and primary ICD-10 diagnosis that led to death for all participants who died during or prior to April 2021. The end of follow-up was defined as death or the end of hospital inpatient data collection in April 2021. Causes of death were grouped according to their ICD codes into malignancies (C00-C97), digestive diseases (K00-K93), cardiovascular diseases (I00-I99), and respiratory diseases (J00-J99).

For the first 50,000 subjects, genotyping was performed using the UK BiLEVE Axiom Array, while the Affymetrix UK Biobank Axiom Array was used for the remaining 450,000 participants. HLA-DQ2.5 status was determined using the SNP rs2187668.

The study was approved by the UKB Access Committee (Project #47527). All participants provided written consent for the study. The UK Biobank study has approval from the Northwest Multicentre Research-Ethics Committee. The manuscript is based solely on the analysis of pseudonymized data obtained from the UK Biobank.
Resource under Application Number 47527. The authors were not in contact with the described individuals, nor had they access to their personal data. We confirm that all research was performed in accordance with the Declaration of Helsinki and other relevant guidelines. The data were reported as described by the STROBE guidelines.

**PheWAS analysis.** Participants were assigned to the disease-specific subgroups based on their ICD-10 diagnoses. Subjects with more than one of the analysed conditions (i.e., CeD, UC, or CD) were excluded (n = 868). The remaining participants were used as the reference cohort (Supplementary Fig. S1). ICD-10 diagnoses obtained from medical reports were collected for each subject, and duplicates were removed. To perform a PheWAS analysis, all ICD-10 codes were converted into Phecodes using the "PheWAS" R package. Phecodes represent a high-throughput phenotyping tool used to rapidly define the case/control status of clinically meaningful diseases and conditions. Using this package, a series of case–control tests were performed: (1) each analysed case group was generated by including patients with the corresponding Phecode; (2) individuals were assigned to the control group when they lacked the tested Phecode; and (3) to ensure statistical power, analysis was restricted to Phecodes with at least 200 cases. Autoimmune Phecodes were identified using the official list of the American autoimmune association.

**Statistical analysis.** All continuous variables are presented as the mean ± standard deviation. Categorical variables are displayed as absolute and relative frequencies. Odds and hazard ratios are presented with their corresponding 95% confidence intervals (CIs). Hazard ratios were calculated using Cox proportional hazard regression models. Mortality was depicted as deaths per 1000 person-years, which were calculated using the following formula: (number of deaths/total number of subjects)/mean survival * 1000. To test for independent associations, multivariable logistic regression was used. All multivariable analyses were adjusted for age, sex and body mass index (BMI). PheWAS analysis was performed using the "PheWAS" R package. Bonferroni correction was used to adjust for multiple testing, and differences were considered to be statistically significant when p < 0.05. The data were analysed using SPSS Statistics version 27 (IBM; Armonk, NY, USA) and visualized with Prism version 8 (GraphPad, La Jolla, CA, USA).

**Data availability**
The data underlying this article are part of the UK Biobank database (https://www.ukbiobank.ac.uk/) and can be accessed after prior application. This research has been conducted under Application Number 47527.

Received: 6 December 2021; Accepted: 16 August 2022
Published online: 26 August 2022

**References**
1. Kaplan, G. G. The global burden of IBD: from 2015 to 2025. *Nat. Rev. Gastroenterol. Hepatol.* **12**(12), 720–727 (2015).
2. Green, P. H. & Cellier, C. Celiac disease. *N. Engl. J. Med.* **357**(17), 1731–1743 (2007).
50. Wu, P., Gifford, A., & Meng, X., et al. Developing and evaluating mappings of ICD-10 and ICD-10-CM codes to phecodes. bioRxiv 2019.
51. Verma, A. et al. A simulation study investigating power estimates in phenome-wide association studies. BMC Bioinf. 19(1), 120 (2018).
52. Association AARD. Autoimmune Disease List. https://www.aarda.org/diseaseList/ (accessed 03.06.2021).

**Author contributions**
Acquisition of data: M.K., C.V.S. Analysis and interpretation of data: M.K., C.V.S., T.B., P.S. Drafting the manuscript: M.K, C.V.S., T.B., P.S. Critical revision of the manuscript for important intellectual content: M.K., C.V.S., T.B., P.S. Figures and tables: M.K., C.V.S. Statistical analysis: M.K. Obtained funding: C.V.S., P.S. Study supervision: P.S. All authors had full access to all of the data and approved the final version of this manuscript. All authors take responsibility for the integrity of the data and the accuracy of the data analysis.

**Funding**
Open Access funding enabled and organized by Projekt DEAL. P.S. is supported by the German Research Foundation (DFG) consortium SFB 1382 “Gut-liver axis” (SFB1382 Project ID 403114013/A03) and DFG grant STR 1095/6–1 (Heisenberg professorship). C.V. S. is supported by Walter-Benjamin Fellowship from the German Research Foundation (SCHN-1640/1–1). T.B. is supported by grants from the German Research Foundation (DFG) (SFB1382 Project ID 403114013/B07). This research was conducted using the UK Biobank Resource under Application Number 47527.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18593-y.

**Correspondence** and requests for materials should be addressed to P.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022