Original research

Phenotyping, genotyping, and prediction of abdominal pain in children using machine learning

Kazuya Takahashi[1,2], Huma Shehwana[3], James K. Ruffle[1,4], John A Williams[5,6,7], Animesh Acharjee[5,6,7], Shuji Terai[2], Georgios V Gkoutos[5,6,7], Humayoon Satti[3], and Qasim Aziz[1]

1. Centre for Neuroscience and Trauma, Wingate Institute of Neurogastroenterology, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK.

2. Division of Gastroenterology and Hepatology, Graduate School of Medical and Dental Sciences, Niigata University, Niigata, Japan.

3. Department of Biological Sciences, National University of Medical Sciences, Rawalpindi, Pakistan.

4. Queen Square Institute of Neurology, University College London, UK

5. College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK.

6. Health Data Research UK, Midlands Site, Birmingham, UK.

7. Centre for Health Data Science, Birmingham, UK

H.Satti and Q.Aziz are joint senior authors.

Corresponding Author: Qasim Aziz

Centre for Neuroscience and Trauma, Wingate Institute of Neurogastroenterology, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Tel: +44 2078822655

E-mail: q.aziz@qmul.ac.uk

Abstract

Objective

Mechanisms of abdominal pain in children are not fully understood due to patient heterogeneity. We aimed to identify abdominal pain phenotypes in children to facilitate the investigation of phenotypic-genotypic associations and to determine risk factors for abdominal pain.

Design

This study included 13,789 children from a large birth cohort. The comorbidities of children and mothers and single nucleotide polymorphisms in children were investigated. Machine learning (ML) was used to identify clusters of patients with homogenous characteristics; subsequently, genome-wide association studies and enrichment analyses were performed. The factors contributing to predictive models were identified using ML.

Results

A total of 1,274 children experienced abdominal pain (9.2 %) (average age: 8.4 ± 1.1 years old, male/female: 615/659), who were classified into 3 clusters: cluster 1 with an allergic predisposition (n = 137), cluster 2 with mother's comorbidities (n = 676), and cluster 3 with minimal comorbidities (n = 340). Enrichment analysis indicated that

genetic pathways related to the intestinal barrier or bile acid biosynthesis were associated with abdominal pain in cluster 1; bile acid biosynthesis was also involved in cluster 2. Predictive models demonstrated modest fidelity with AUC values up to 0.65 in predicting children's abdominal pain, showing mother's and children's comorbidities formed significant risk factors.

Conclusion

The risk factors and phenotypes of paediatric abdominal pain are embedded within phenotype-genotype associations, which can be targeted in future studies. In particular, the link between allergy and intestinal barrier may be of mechanistic and therapeutic importance.

Keywords: abdominal pain, children, machine learning, artificial intelligence, Uniform Manifold Approximation and Projection, Genome-wide association studies, enrichment analysis

What is already known on this topic

- Although abdominal pain in children is rarely life-threatening in children, it is however a very common symptom leading to impaired quality of life.

- Abdominal pain in children is heterogeneous and has various associated factors such as gastrointestinal, psychological, and allergic diseases, diseases causing somatic pain, and parents' comorbidities. Robust stratification is, therefore, necessary to elucidate its etiology and pathophysiology.

- Machine learning can stratify complicated multidimensional datasets in medical and biomedical fields by identifying disease clusters with pathophysiological implications and is a powerful tool that aids in predicting the development of diseases or treatment outcomes.

What this study adds

- There were 3 clusters of paediatric abdominal pain governed by the following: 1) those with allergic predispositions, 2) those with high maternal comorbidities, and 3) those with minimal paediatric and maternal comorbidities.

- Genome-wide association studies and enrichment analyses revealed that different genes associated with biological processes and pathways might be responsible for

phenotypic differences between paediatric pain clusters. The link between allergic diseases in cluster 1 and the intestinal barrier may be of mechanistic and therapeutic importance.

- Machine learning models predicting abdominal pain demonstrated moderate performance and indicated that mother's comorbidities as well as children's clinical and genetic information formed significant risk factors.

How this study might affect research, practice or policy

- Each phenotype of abdominal pain in children may need a tailor-made therapeutic approach according to its etiology and pathophysiology.

- Significant pathways in the enrichment analysis, such as those implicating intestinal barrier permeability, can be a target for future studies to investigate the mechanisms of abdominal pain.

- Machine learning models, coupled with further relevant information, such as socioeconomic and biochemical factors, may be able to precisely predict the development of children's abdominal pain in clinical settings.

Introduction

Abdominal pain is one of the most common symptoms among children and adolescents, with prevalence rates across the USA and Europe ranging from 0.3% to 19.0% [1–3]. In primary care, most children presenting with abdominal pain are diagnosed with functional abdominal pain (FAP), whereas organic causes account for only 5% to 10% of cases [4,5]. Whilst abdominal pain is not life-threatening in most paediatric cases, it is often refractory to treatment and associated with psychiatric comorbidities such as anxiety and depressive disorders [6–8]. Abdominal pain in children has a significant impact on their health-related quality of life [9], and hence, early recognition and intervention are warranted.

Organic diseases, as well as early life events ranging from psychological abuse, allergy status, psychological comorbidity, and parental factors, are proposed risk factors for abdominal pain in children [5,10–15]. However, not all children with these risk factors develop abdominal pain, leaving the mechanisms of paediatric abdominal pain

incompletely understood. Moreover, to date, while there have been only a few reports of genetic associations for abdominal pain in children, genetic variation in the NPSR1 gene, which codes for the neuropeptide S receptor, have being reported to influence children's predisposition to recurrent abdominal pain [16]. However, abdominal pain was not well phenotyped in this study and could therefore account for various causes. To better understand the determinants of abdominal pain and provide novel insights into phenotypic-genotypic associations of abdominal pain in children, a robust stratification in a large sample of patients is necessary.

One such stratification method is a data-driven clustering by machine learning (ML) that can visualize and stratify complicated multidimensional datasets to identify disease clusters with pathophysiological implications [17–20]. Furthermore, ML based modelling is a powerful tool that may aid clinicians to effectively detect the risk of abdominal pain development in early childhood [21]. We hypothesised that a ML algorithm could stratify children with abdominal pain into clusters according to common characteristics. Moreover, in elucidating phenotypes of abdominal pain this will plausibly help to identify phenotype-genotype associations. We further hypothesised that ML algorithms could help quantify the important factors that predict

the development of abdominal pain.

Herein, we systematically quantify the risk factors of paediatric abdominal pain in a large birth cohort. In applying unsupervised ML, we delineate phenotypic pain association criterion to not only the child, but also the mother. We extended gene analysis to these clusters to delineate specific pathway associations. Finally, ML models were used to predict the development of paediatric abdominal pain with modest fidelity when utilizing information from both the child and the mother.

Methods

Study subjects

Between 2007 and 2011, 12,453 pregnant women (recruited at 26 - 28 weeks) and 13,858 children were registered in the Born in Bradford (BiB) cohort. The demographics of the whole cohort are described elsewhere [22]. Linked general practitioner (GP) records were available for 13,789 of these children, and all mothers. The study population is described in Figure 1. We first investigated the abdominal pain prevalence associated with any cause using systematized nomenclature of medicine-clinical terms (SNOMED-CT). Abdominal pain was defined when patients had one of

the following diagnoses included in SNOMED-CT regardless of being acute or chronic: abdominal pain, abdominal wall pain, or generalized abdominal pain. The frequencies of comorbidities, such as gastrointestinal (GI), psychological, and allergic diseases and diseases causing somatic pain, were also investigated using SNOMED-CT since these are associated with organic and functional abdominal pain [10,12–14,23–25]. Furthermore, the comorbidities of children's mothers were investigated to determine the effect of mothers' comorbidities on the development of abdominal pain in children. The diseases of children and their mothers investigated in this study are listed in supplementary material 1. The demographics and the frequencies of diseases were compared between children with and without abdominal pain. Fathers were not included in our analysis pipeline, as there were a significant proportion of missing values (in excess of 74.7 %).

Statistical analysis

Categorical data were expressed as numbers plus percentages. Continuous data were expressed as mean ± standard deviation (SD). The Student’s t-test and Fisher's exact test were used for numerical data and categorical data, respectively. The Cox proportional hazards model was used to evaluate risk factors for abdominal pain in 13,789 children.

A p-value < 0.05 was considered statistically significant. EZR (Saitama Medical Center, Jichi Medical University, Saitama, Japan) was used for statistical analysis [26].

Unsupervised clustering

We performed clustering of 1,274 children with abdominal pain by ML, using Python (version 3.7.12) [27]. Uniform Manifold Approximation and Projection (UMAP), a non-linear dimension reduction technique was applied to 27 features, reducing them to 3 dimensions for later clustering [28]. These non-linear embeddings were then used for clustering with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [29]**.** Scatterplots were produced to visualise the characteristics of each cluster. GI, psychiatric, somatic pain, and allergic comorbidities of children and their mothers were used for the clustering, which were summarised in supplemental material 2. Clinical characteristics of children and their mothers were compared among clusters detected to derive broad feature-weights for their allocation. Adjusted p-values were calculated by multiplying p-values with the number of comparisons (Bonferroni correction) when multiple comparisons were performed to compare the clinical characteristics of each cluster. Adjusted p-value < 0.05 was considered statistically significant.

Genome-wide association studies (GWAS)

We performed GWAS based on the clustering results to seek significant variants of single nucleotide polymorphisms (SNPs) responsible for the differences among clusters. The quality control process is summarized in supplemental material 3. SNPs data were available in 64 children in cluster 1, 346 in cluster 2, and 160 in cluster 3. Using 6,240 children without abdominal pain who had SNPs data as the control, case-control association tests were performed between each cluster and the control with an additive logistic regression model implemented in PLINK [30]. Manhattan plots were made based on the results (supplemental material 4). For performing multiple comparisons in the GWAS data analysis, false discovery rate (FDR) adjusted *P*-values were also calculated using Benjamini-Hochberg correction with a threshold for statistical significance $p < 0.05$.

Functional and pathway enrichment analysis

The top 10 SNPs, with an adjusted p-value < 0.05 in GWAS, were selected in each cluster, and the gene annotations for each SNP were obtained from the HaploReg online portal (http://pubs.broadinstitute.org/mammals/haploreg/haploreg.php), making gene sets for each cluster. Then, the gene sets were analysed by Enrichr

(https://maayanlab.cloud/Enrichr/) for the gene function annotation and the identification of their associated signaling pathway [31]. Gene ontology (GO) terms were used for functional enrichment analysis, investigating biological process (BP), cellular composition (CC), and molecular function (MF) of the gene sets. The biological pathways associated with the gene sets were investigated by pathway enrichment analysis with the Kyoto Encyclopedia of Genes and Genomes (KEGG). An FDR adjusted p-value < 0.05 was considered statistically significant in the enrichment analyses.

Supervised predictive models

We then developed four predictive models to identify the development of abdominal pain in children and used the extreme gradient-boosted tree (XGB) classifier, a widely used supervised machine learning algorithm with high effectiveness described in detail elsewhere [32]. Data were partitioned randomly to 80% for model training and 20% for out-of-sample testing. We sought to evaluate the additive value of feature categories, constructing four separate models as follows. Firstly, we trained a model with all clinical features and SNPs data (Model 1); secondly, we trained a model using only clinical features (Model 2); thirdly we trained a model using only SNPs (Model 3).

Since large numbers of features can result in poor out-of-sample generalisation [33], we additionally trained a final fourth model using 20 features with a feature reduction (Model 4) [33,34]. Thirty SNPs, listed in Table 4, were used for the models. The features used for predictive models and the detailed setting of the hyper-parameters are summarized in supplemental material 5.

The performances of the predictive models were assessed against the test dataset, using the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, precision, recall, and F1 score. We used Shapley Additive exPlanations (SHAP) value to evaluate how each feature contributed to the models [35,36].

Results

Among 13,789 children, 1,274 children (9.2%) experienced abdominal pain (male/female = 615/659). The average age of children when the data was extracted was 8.4 ± 1.1 years old. The average age when children were diagnosed with abdominal pain was 5.6 ± 2.7 years old.

Comparison between 1,274 children with abdominal pain and 12,515 children without abdominal pain (Table 1)

There were more females in the abdominal pain group than in the non-abdominal pain group (51.7 % vs. 48.1 %, $p = 0.01$). Furthermore, more children were of south Asian origin in the abdominal pain group compared to the non-abdominal pain group, such as Bangladeshi (5.7 %, vs. 2.9 %, $p < 0.01$) and Pakistani (58.6 % vs. 45.8 %, $p < 0.01$), and fewer children of white British origin in the abdominal pain group (22.3 % vs. 36.4 %, $p < 0.01$). In children with abdominal pain, comorbidities were more prevalent; these included allergic diseases (47.2 % vs. 37.3 %, $p < 0.01$), constipation (4.5 % vs. 1.1 %, $p < 0.01$), and gastro-oesophageal reflux disease (GORD) (5.2 % vs. 2.7 %, $p < 0.01$), compared to those without abdominal pain.

Mothers of children with abdominal pain compared to those without reported themselves experiencing more abdominal pain (49.5 % vs. 31.8 %, $p < 0.01$), allergic diseases (47.8 % vs. 39.4 %, $p < 0.01$), arthritis (8.6 % vs. 4.0 %, $p < 0.01$), depressive and/or bipolar disorder (24.3 % vs. 21.0 %, $p = 0.01$), GORD (13.4 % vs. 7.0 %, $p < 0.01$), irritable bowel syndrome (IBS) (9.9 % vs. 7.0 %, $p < 0.01$), and migraine (22.4 % vs. 14.7 %, $p < 0.01$).

Risk factors for abdominal pain

The significant results of the Cox proportional hazards model are summarised in Table 2, and non-significant results are further detailed within supplemental material 6. Female gender (HR [95% CI]: 1.1 [1.0–1.3], $p = 0.03$) was a statistically significant risk factor for abdominal pain. Ethnicities, such as Bangladeshi (2.6 [2.0-3.4], $p < 0.01$), Indian (2.1 [1.5-2.9], $p < 0.01$), and Pakistani (1.8 [1.6-2.1], $p < 0.01$), were also significant risk factors (when compared to. White British). Within children, comorbidities such as allergic disease (1.1 [1.0-1.3], $p = 0.04$), appendicitis (5.1 [3.4-7.6], $p < 0.01$), Celiac disease (3.0 [1.5-5.8], $p < 0.01$), constipation (2.5 [1.9-3.3], $p < 0.01$), functional dyspepsia (8.2 [3.0-22.4], $p < 0.01$), IBS (5.1 [2.3-11.5], $p < 0.01$), and migraine (2.8 [1.7-4.5], $p < 0.01$) were significant risk factors for abdominal pain. Within mother's comorbidities, abdominal pain (1.6 [1.4-1.8], $p < 0.01$), allergic disease (1.2 [1.0-1.3], $p = 0.02$), arthritis (1.4 [1.2-1.8], $p < 0.01$), GORD (1.4 [1.2-1.7], $p < 0.01$), and migraine (1.2 [1.1-1.4], $p = 0.01$) were significant risk factors for paediatric abdominal pain.

Phenotyping and genotyping of children with abdominal pain by unsupervised ML clustering

The unsupervised model classified children with abdominal pain into 3 unique clusters: 137 children in cluster 1 (10.8 %), 677 children in cluster 2 (53.1 %), and 340 children in cluster 3 (26.7 %) (Figure 2). The remaining 120 patients (9.4 %) were not classified into any clusters because they had miscellaneous characteristics and were scattered sparsely on the plot; these children were excluded from downstream analyses.

The significant and non-significant differences in clinical characteristics among 3 clusters are summarised in Table 3 and supplemental material 7, respectively. Furthermore, the top 10 significant SNPs of each cluster, detected by GWAS, as well as the annotated genes, are listed in Table 4. The results of enrichment analysis based on the annotated genes are described in Table 5 (the visualization of the results was summarised in supplemental material 8). The summary of clinical and genetic characteristics of each cluster is described below.

Cluster 1 (Cluster with allergic predisposition)

Most children had allergic diseases (99.3 %), and all mothers had allergic diseases (100 %), which were significantly higher than in clusters 2 and 3 ($p < 0.01$). On the

other hand, mother's abdominal pain was not observed in this cluster. In short, this cluster was characterized as a ‘cluster with allergic predisposition’.

The pathway enrichment analysis revealed that both actin related protein 2/3 complex subunit 5 (ARPC5) (adjusted p-value = 0.03) and CLDN16 (adjusted p-value = 0.03) were significantly enriched in the pathways, ‘Tight junction’ and ‘Pathogenic Escherichia coli infection’ (Table 5). Also, cytochrome P450 family 7 subfamily B member 1 (CYP7B1) was significantly enriched in ‘Primary bile acid biosynthesis’ (adjusted p-value = 0.04).

Cluster 2 (Cluster with mother's comorbidities)

Mothers' comorbidities frequencies, such as abdominal pain (70.0 %), allergic (40.0 %), depressive and/or bipolar disorder (37.2 %), GORD (18.9 %), migraine (34.0 %), were the highest in this cluster compared to the other 2 clusters. We termed this as a ‘cluster with mother's comorbidities.’ The frequency of children’s allergic disease was 68.7 %, albeit less than the frequency identified in cluster 1.

Functional enrichment analysis revealed various significant GO terms, mainly related to the regulation of extrinsic apoptotic signaling and bile acid biosynthesis in BP. Furthermore, oxysterol-binding protein-related protein 3 (OSBPL3) was enriched in

‘cholesterol binding’ (adjusted p-value = 0.03) and ‘sterol binding’ (adjusted p-value = 0.03), and EYA transcriptional coactivator and phosphatase 4 (EYA4) was enriched in in 'protein tyrosine phosphatase activity' (adjusted p-value = 0.03) in MF. There were no significantly enriched genes in CC and pathway enrichment analysis.

Cluster 3 (Cluster with the least comorbidities)

The frequency of mother's abdominal pain (46.2 %) was the second highest among the clusters. However, the frequencies of other comorbidities, including allergic diseases and mother's comorbidities, were very low within this cluster. This cluster was termed as a ‘cluster with the least comorbidities’.

No significant terms were identified in the enrichment analyses of this cluster.

Prediction of children with abdominal pain vs. those without

Model 1, a supervised machine classifier predicted the presence of child abdominal pain using paediatric clinical and SNP data and maternal clinical data, showed the highest performance among the four models, with the following test set performance: 0.65, accuracy: 0.91, precision: 0.38, recall: 0.11, F1 score: 0.17. Meanwhile the performances of Model 2, utilizing only paediatric and maternal *clinical* features,

achieved a weaker fidelity with the following test set performance: AUC: 0.55, accuracy: 0.77, precision: 0.10, recall: 0.21, F1 score: 0.14. Model 3, incorporating only paediatric SNPs data showed further reduced fidelity with test performances of AUC: 0.57, accuracy: 0.86, precision: 0.10, recall: 0.08, F1 score: 0.09. Model 4, using only with 20 selected features achieved a performance superior to Model 2 and 3, but inferior to Model 1, with the following test set metrics: 0.59, accuracy: 0.83, precision: 0.13, recall: 0.18, F1 score: 0.15). The results of Model 1 and the other models are demonstrated in Figure 3 and supplemental material 9, respectively.

In reviewing SHAP feature importance for the best performing model (Model 1), the greatest determinant of a child developing abdominal pain was the presence of abdominal pain in the mother (Figure 3). Additional clinicodemographic determinants of child abdominal pain were if the child were female, had a history of allergic disease, and ethnicity other than non-white British, in our sample; 12 of the top 20 determinants of child abdominal pain were also related to child genetic data (Figure 3, panel b-c). Furthermore, a maternal history of migraine, allergic disease, depression and/or bipolar disorder were additional factors featuring in the top 20 determinants of child abdominal pain.

Discussion

Summary of findings

The frequency of children's abdominal pain in the study cohort was 9.2 %. The Cox proportional hazards model demonstrated some paediatric and maternal risk factors for abdominal pain. ML-based clustering successfully identified three phenotypes of abdominal pain in children. GWAS performed against the clustering results, and subsequent enrichment analysis indicated different mechanisms of abdominal pain in each cluster. The performance of the predictive model for abdominal pain were moderate with the information from GP records and SNPs data, where the greatest determinants of paediatric abdominal pain were maternal comorbidities (especially abdominal pain), paediatric gender, ethnicities, allergic diseases, and SNP data.

Risk factors for abdominal pain in children

Our study showed that mother's comorbidities were significant risk factors for paediatric abdominal pain. Previous studies have shown that parental factors are associated with FAP in children [10,13,14]. For example, children of mothers with IBS

had more GI symptoms, and parents' solicitous response to illness affected children's symptoms [14]. Moreover, mothers of children with FAP tended to have more IBS and migraine [37]. Thus, mothers' comorbidities and children's abdominal pain are associated with each other. Our results were concordant with the previous reports.

Our study also revealed that allergic predispositions were a risk factor for paediatric abdominal pain. Some studies have reported that allergic diseases form significant risk factors for functional GI disorders such as IBS [11,24,38]. Furthermore, a recent study revealed that injection of food antigens into the rectosigmoid colon of IBS patients induces local edema and mast cell activation, proposing IBS may be a food-induced disorder mediated by mast cell activation localized to the intestine [39]. This previous study implies that local immune reactions are one of the possible causes of abdominal pain in children with allergic predispositions.

Ethnicity was also a significant risk factor for abdominal pain in this study. The BiB cohort is unique in that children are mainly comprised of White British (29.3%) and Pakistani ethnicities (39.5%) [22]. Socioeconomic status was significantly different among ethnicities in the BiB cohort [40], and lower socioeconomic status is a risk factor

for pain conditions in children, including abdominal pain [41]. Therefore, the variations in the cohort's socioeconomic status might be relevant to the abdominal pain variations.

Cluster analysis identifies 3 distinct phenotypes and genotypes of abdominal pain in children

The ML clustering revealed 3 distinct phenotypes of abdominal pain in children. The aforementioned allergic mechanism could account for abdominal pain in cluster 1 [39]. Furthermore, the enrichment analysis showed some molecular functions and pathways that might explain abdominal pain experienced by the cluster 1 participants. ARPC5 and CLDN16 were significant genes related to the pathway, 'Tight junction', which may explain the characteristics of cluster 1. The intestinal epithelial barrier consists of interepithelial protein complexes, including tight junctions, adherens junctions, and the actin cytoskeleton [42]. ARPC5 protein is an important actin regulator responsible for actin cytoskeletal remodeling associated with compromised intestinal epithelium [42]. Furthermore, claudins are major structural components of tight junctions of the intestinal epithelium. Hence, SNPs associated with these genes may result in barrier dysfunction of the intestine. Interestingly, barrier dysfunction of the GI tract is associated with abdominal pain, and increased permeability of the epithelial cells is

associated with allergic diseases such as atopic eczema and bronchial asthma [43–46], which are consistent with allergic predispositions of children in cluster 1. Furthermore, ARPC5 and CLDN16 were also enriched in the pathway ‘Pathogenic E.coli infection’, which also seemed important for abdominal pain mechanism in this cluster. Enteropathogenic *Escherichia coli* (EPEC) is a pathogen causing paediatric diarrhea [47]. Barrier dysfunction of the GI tract may result in vulnerability against EPEC infection in children. Infectious gastroenteritis by *E.coli* is a well-known risk factor for subsequent IBS in children, causing abdominal pain [48,49]. Therefore, abdominal pain in this cluster might be partly explained by the effect of infectious gastroenteritis by EPEC.

‘Primary bile acid biosynthesis’ is also an important pathway for abdominal pain in cluster 1, in which CYP7B1 was enriched [50]. In patients with IBS, the metabolism of primary bile acids, which is influenced by microbiota, is associated with diarrhea and abdominal pain intensity [51]. In addition, colonic bile acid membrane receptor expression correlates with abdominal pain severity [52]. Therefore, malfunction of primary bile acid biosynthesis may affect paediatric abdominal pain.

Thus, ARPC5, CLDN16, and CYP7B1 could be responsible for abdominal pain in cluster 1. Our findings are novel given that there have previously been no reports investigating the association between paediatric abdominal pain and these genes regulating intestinal permeability and bile acid synthesis, which may be because of a lack of identification of specific patient clusters where phenotype–genotype correlation is more meaningful.

In cluster 2, mother's comorbidities were more frequent than in other clusters. It is generally accepted that children's abdominal pain is related to their parent’s illness behaviour, which may explain the mechanism of abdominal pain in cluster 2 [53]. Furthermore, enrichment analysis showed various significant GO terms, although it remains unknown how these biological processes, such as apoptosis or DNA repair, etc., are associated with abdominal pain. However, terms related to cholesterol and bile acid metabolism/ biosynthesis were significant in this cluster, which is a possible underlying abdominal pain mechanism in cluster 2 [50] since similar terms were significant in the analysis of cluster 1. Processes related to bile acid metabolism and synthesis may cause abdominal pain in a wide range of children.

In contrast, children in cluster 3 are unique in that they rarely had comorbidities. Although mother's abdominal pain was relatively frequent (46.2%), children in this cluster had abdominal pain that allergic or other parental factors could not explain, and the enrichment analysis revealed no significant results. Like clusters 1 and 2, the frequencies of organic GI diseases responsible for acute abdominal pain were low in this cluster. Therefore, factors, such as socioeconomic status, might exist, and information available from the GP records was possibly insufficient to deeply phenotype children in this cluster, resulting in negative results in the subsequent genetic analysis.

Machine learning predictive models predicted abdominal pain in children with moderate performance

Model 1 showed moderate predictive performance for abdominal pain in children. The combination of both the paediatric and maternal clinical data and the genetic data yielded the best predictive performance. This result implies that the development of abdominal pain in children is potentially a result of an interacting combination of their own innate factors and that of their surrounding family/genetics. Among features used in the best model (Model 1), mothers' comorbidities, especially abdominal pain, were

more significant predictors than children's own healthcare and genetic data. This result implies that mother's illness behaviour may have a stronger impact on the development of paediatric abdominal pain than other factors [53]. Epigenetic alteration, changes in microbiota, and dysfunction of the autonomic nervous system could also affect abdominal pain [54–56]. Therefore, further information related to these factors may improve the performance of the predictive models.

Limitations

Our study has some limitations. Firstly, the diagnosis of diseases was based on GP records; therefore, the criteria used for each diagnosis and the exact causes of abdominal pain in each case are unknown. However, since most children with abdominal pain are diagnosed with FAP at the primary care level [5], and the frequencies of organic diseases related to abdominal pain were rare in this cohort, FAP would be the most likely cause of abdominal pain. Second, although we evaluated the performance of our predictive models by a standard method, we did not validate them using a different cohort of patients to evaluate the generalisability of the models. Therefore, although rigorously assessed, model overfitting is plausible.

Conclusions

We report that children's and mother's comorbidities form significant risk factors for abdominal pain in children. The novel data-driven clustering detected 3 distinct phenotypes of abdominal pain in children: abdominal pain with an allergic predisposition, mother's comorbidities, and the least comorbidities, which were useful for seeking phenotype-genotype association. Each phenotype may need a tailor-made therapeutic approach according to its etiology and pathophysiology. Furthermore, findings detected in the enrichment analysis can be a target for future studies to investigate the mechanisms of abdominal pain. Further analysis investigating relevant socioeconomic and biochemical factors, combined with our study's ML-based approach, may enable more precise phenotyping and prediction.

Funding

This study was supported by JA Niigata Kouseiren Grant (Niigata University School of Medicine).

References

1. Chitkara DK, Rawat DJ, Talley NJ. The epidemiology of childhood recurrent abdominal pain in western countries: A systematic review. Am J Gastroenterol. 2005;100: 1868–1875.

2. Korterink JJ, Diederen K, Benninga MA, Tabbers MM. Epidemiology of pediatric functional abdominal pain disorders: A meta-analysis. PLoS One. 2015;10: 1–17.

3. Devanarayana NM, de Silva DGH, de Silva HJ. Recurrent abdominal pain syndrome in a cohort of Sri Lankan children and adolescents. J Trop Pediatr. 2008;54: 178–183.

4. Vila M, Kramer T, Obiols JE, Garralda ME. Abdominal pain in British young people: Associations, impairment and health care use. Journal of Psychosomatic Research. 2012. pp. 437–442. doi:10.1016/j.jpsychores.2012.09.009

5. Spee LA, Van Den Hurk AP, Van Leeuwen Y, Benninga MA, Bierma-Zeinstra SM, Passchier J, et al. Childhood abdominal pain in primary care: Design and patient selection of the HONEUR abdominal pain cohort. BMC Fam Pract. 2010;11: 6–8.

6. Dufton LM, Dunn MJ, Compas BE. Anxiety and somatic complaints in children with recurrent abdominal pain and anxiety disorders. J Pediatr Psychol. 2009;34: 176–186.

7. Garralda ME. Unexplained physical complaints. Pediatr Clin North Am. 2011;58: 803–813.

8. Chu A, Torres L, Kao G, Gilbert C, Monico EC, Chumpitazi BP. Multidisciplinary Care for Refractory Pediatric Functional Abdominal Pain Decreases Emergency and Inpatient Utilization. J Pediatr Gastroenterol Nutr. 2021;Publish Ah. doi:10.1097/mpg.0000000000003305

9. Warschburger P, Hänig J, Friedt M, Posovszky C, Schier M, Calvano C. Health-related quality of life in children with abdominal pain due to functional or organic gastrointestinal disorders. J Pediatr Psychol. 2014;39: 45–54.

10. Korterink J, Devanarayana NM, Rajindrajith S, Vlieger A, Benninga MA. Childhood functional abdominal pain: Mechanisms and management. Nat Rev Gastroenterol Hepatol. 2015;12: 159–171.

11. Sjölund J, Kull I, Bergström A, Järås J, Ludvigsson JF, Törnblom H, et al. Allergy-related diseases in childhood and risk for abdominal pain-related functional gastrointestinal disorders at 16 years—a birth cohort study. BMC Med. 2021;19: 1–11.

12. McOmber M, Rafati D, Cain K, Devaraj S, Weidler EM, Heitkemper M, et al.

Increased Gut Permeability in First-degree Relatives of Children with Irritable Bowel Syndrome or Functional Abdominal Pain. Clin Gastroenterol Hepatol. 2020;18: 375-384.e1.

13. Zeevenhooven J, Rutten JMTM, van Dijk M, Peeters B, Benninga MA. Parental Factors in Pediatric Functional Abdominal Pain Disorders: A Cross-sectional Cohort Study. J Pediatr Gastroenterol Nutr. 2019;68: e20–e26.

14. Levy RL, Whitehead WE, Walker LS, Von Korff M, Feld AD, Garner M, et al. Increased somatic complaints and health-care utilization in children: Effects of parent IBS status and parent response to gastrointestinal symptoms. Am J Gastroenterol. 2004;99: 2442–2451.

15. Nimnuan C, Hotopf M, Wessely S. Medically unexplained symptoms: an epidemiological study in seven specialities. J Psychosom Res. 2001;51: 361–367.

16. Henström M, Zucchelli M, Söderhäll C, Bergström A, Kere J, Melén E, et al. NPSR1 polymorphisms influence recurrent abdominal pain in children: A population-based study. Neurogastroenterol Motil. 2014;26: 1417–1425.

17. Ruffle JK, Farmer AD, Aziz Q. Artificial Intelligence-Assisted Gastroenterology — Promises and Pitfalls. Am J Gastroenterol. 2019;114: 422–428.

18. Yang Y, Sun H, Zhang Y, Zhang T, Gong J, Wei Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep. 2021;36: 109442.

19. Ko S, Pareek A, Jo C, Han H-S, Lee MC, Krych AJ, et al. Automated Risk Stratification of Hip Osteoarthritis Development in Patients With Femoroacetabular Impingement Using an Unsupervised Clustering Algorithm: A Study From the Rochester Epidemiology Project. Orthop J Sports Med. 2021;9: 23259671211050612.

20. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37: 38–47.

21. Bravo-Merodio L, Acharjee A, Russ D, Bisht V, Williams JA, Tsaprouni LG, et al. Translational biomarkers in the era of precision medicine. Adv Clin Chem. 2021;102: 191–232.

22. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort profile: The born in bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42: 978–991.

23. Spee LAA, Lisman-Van Leeuwen Y, Benninga MA, Bierma-Zeinstra SMA, Berger MY. Prevalence, characteristics, and management of childhood functional abdominal pain in general practice. Scand J Prim Health Care. 2013;31: 197–202.

24. Olén O, Neuman A, Koopmann B, Ludvigsson JF, Ballardini N, Westman M, et al. Allergy-related diseases and recurrent abdominal pain during childhood - A birth cohort study. Aliment Pharmacol Ther. 2014;40: 1349–1358.

25. Fang ZY, Zhang HT, Lu C, Lu QM, Yu CH, Wang HY. Association between Allergic Diseases and Irritable Bowel Syndrome: A Retrospective Study. Int Arch Allergy Immunol. 2018;177: 153–159.

26. Kanda Y. Investigation of the freely available easy-to-use software “EZR” for medical statistics. Bone Marrow Transplant. 2013;48: 452–458.

27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12: 2825–2830.

28. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML]. 2018. Available: http://arxiv.org/abs/1802.03426

29. Melvin RL, Xiao J, Godwin RC, Berenhaut KS, Salsbury FR. Visualizing correlated motion with HDBSCAN clustering. Protein Sci. 2018;27: 62–75.

30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

31. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90-7.

32. Cappon G, Facchinetti A, Sparacino G, Georgiou P, Herrero P. Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes—an in silico proof-of-concept. Sensors . 2019;19: 1–11.

33. Abpeikar S, Ghatee M, Foresti GL, Micheloni C. Adaptive neural tree exploiting expert nodes to classify high-dimensional data. Neural Netw. 2020;124: 20–38.

34. Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in Neuroimaging. Neuroinformatics. 2014;12: 229–244.

35. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv

[cs.AI]. 2017. Available: http://arxiv.org/abs/1705.07874

36. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1802.03888

37. Campo JV, Bridge J, Lucas A, Savorelli S, Walker L, Di Lorenzo C, et al. Physical and emotional health of mothers of youth with functional abdominal pain. Arch Pediatr Adolesc Med. 2007;161: 131–137.

38. Jones MP, Walker MM, Ford AC, Talley NJ. The overlap of atopy and functional gastrointestinal disorders among 23 471 patients in primary care. Aliment Pharmacol Ther. 2014;40: 382–391.

39. Aguilera-Lizarraga J, Florens MV, Viola MF, Jain P, Decraecker L, Appeltans I, et al. Local immune response to food antigens drives meal-induced abdominal pain. Nature. 2021;590: 151–156.

40. Petherick ES, Pearce N, Sunyer J, Wright J. Ethnic and socio-economic differences in the prevalence of wheeze, severe wheeze, asthma, eczema and medication usage at 4 years of age: Findings from the Born in Bradford birth cohort. Respir Med. 2016;119: 122–129.

41. Grøholt E-K, Stigum H, Nordhagen R, Köhler L. Recurrent pain in children, socio-economic factors and accumulation in families. Eur J Epidemiol. 2003;18: 965–975.

42. Chánez-Paredes S, Montoya-García A, Castro-Ochoa KF, García-Cordero J, Cedillo-Barrón L, Shibayama M, et al. The Arp2/3 inhibitory protein Arpin is required for intestinal epithelial barrier integrity. Front Cell Dev Biol. 2021;9: 625719.

43. Hanning N, Edwinson AL, Ceuleers H, Peters SA, De Man JG, Hassett LC, et al. Intestinal barrier dysfunction in irritable bowel syndrome: a systematic review. Therap Adv Gastroenterol. 2021;14: 1756284821993586.

44. Zhang Q, Cheng L, Wang J, Hao M, Che H. Antibiotic-Induced Gut Microbiota Dysbiosis Damages the Intestinal Barrier, Increasing Food Allergy in Adult Mice. Nutrients. 2021;13. doi:10.3390/nu13103315

45. Yamani A, Wu D, Ahrens R, Waggoner L, Noah TK, Garcia-Hernandez V, et al. Dysregulation of intestinal epithelial CFTR-dependent Cl- ion transport and paracellular barrier function drives gastrointestinal symptoms of food-induced anaphylaxis in mice. Mucosal Immunol. 2021;14: 135–143.

46. Niewiem M, Grzybowska-Chlebowczyk U. Intestinal Barrier Permeability in

Allergic Diseases. Nutrients. 2022;14. doi:10.3390/nu14091893

47. Gaytán MO, Martínez-Santos VI, Soto E, González-Pedrajo B. Type Three Secretion System in Attaching and Effacing Pathogens. Front Cell Infect Microbiol. 2016;6: 129.

48. Thabane M, Simunovic M, Akhtar-Danesh N, Garg AX, Clark WF, Collins SM, et al. An outbreak of acute bacterial gastroenteritis is associated with an increased incidence of irritable bowel syndrome in children. Am J Gastroenterol. 2010;105: 933–939.

49. Marshall JK, Thabane M, Garg AX, Clark WF, Salvadori M, Collins SM, et al. Incidence and epidemiology of irritable bowel syndrome after a large waterborne outbreak of bacterial dysentery. Gastroenterology. 2006;131: 445–50; quiz 660.

50. Farrugia A, Arasaradnam R. Bile acid diarrhoea: pathophysiology, diagnosis and management. Frontline Gastroenterol. 2021;12: 500–507.

51. Dior M, Delagrèverie H, Duboc H, Jouet P, Coffin B, Brot L, et al. Interplay between bile acid metabolism and microbiota in irritable bowel syndrome. Neurogastroenterol Motil. 2016;28: 1330–1340.

52. Wei W, Wang H, Zhang Y, Zhang Y, Niu B, Chen S, et al. Faecal bile acids and colonic bile acid membrane receptor correlate with symptom severity of diarrhoea-predominant irritable bowel syndrome: A pilot study. Dig Liver Dis. 2021;53: 1120–1127.

53. Levy RL. Exploring the intergenerational transmission of illness behavior: From observations to experimental intervention. Annals of Behavioral Medicine. 2011. pp. 174–182.

54. Zhou Q, Verne GN. Epigenetic modulation of visceral nociception. Neurogastroenterol Motil. 2022;34: e14443.

55. Gibler RC, Jastrowski Mano KE. Systematic Review of Autonomic Nervous System Functioning in Pediatric Chronic Pain. Clin J Pain. 2021;37: 281–294.

56. Rajindrajith S, Zeevenhooven J, Devanarayana NM, Perera BJC, Benninga MA. Functional abdominal pain disorders in children. Expert Rev Gastroenterol Hepatol. 2018;12: 369–390.

Figure legends

Figure 1. Study population

Among 13,789 children from the Born in Bradford cohort, 1,274 children with abdominal pain were included in the clustering analysis by the machine learning algorithm. Meanwhile, 5,876 children with clinical and genetic data were included to make predictive models for abdominal pain.

Figure 2. Clustering of 1,274 children with abdominal pain

(a) Scattering plots of the twenty-seven features reduced to 3 dimensions by unsupervised uniform manifold approximation and projection (UMAP). Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) found 3 distinct clusters. (b) Scattering plots labeled by representative features of children and their mothers.

Figure 3. Result and SHAP feature importance (top 20) of Model 1

(a) Receiver operating characteristics (ROC) curve of Model 1. The areas under the curve (AUC) for Model 1 is 0.65. (b) The bar graph shows the mean SHAP value of each feature. The mother's abdominal pain contributed the most to Model 1. Among paediatric clinical factors, gender, ethnicities, and allergic diseases contributed to Model

1. Moreover, 12 SNPs data were within the top 20 important contributors. (c) The beeswarm plot showed the SHAP feature importance and the direction of the effect of each feature on Model 1. The horizontal axis represents the SHAP value. Since all the features used in the predictive models were binary, red and blue plots showed positive or negative for each feature, respectively. For example, taking a look at the 'Mother's abdominal pain', red plots tend to distribute in the positive SHAP value and blue pots in the negative SHAP value. This means that children with mother's abdominal pain are more likely to have abdominal pain compared to those without.

Table 1. Background of 13,789 children and their frequency of diseases

Background	**With abdominal pain (n = 1274)**	**Without abdominal pain (n = 12515)**	**P value**
Average age at the extraction ± SD (years)	8.4 ± 1.1	8.3 ± 1.1	<0.01
Average age at the diagnosis of abdominal pain ± SD (years)	5.6 ± 2.7	N/A	N/A
Gender, female (%)	659 (51.7 %)	6015 (48.1 %)	0.01
Route of birth, (vaginal/ caesarean)*	943 (75.3 %)	9415 (77.0 %)	0.17
Ethnicity**			
Bangladeshi	68 (5.7 %)	304 (2.9 %)	<0.01
Indian	41 (3.4 %)	296 (2.9 %)	0.28
Pakistani	703 (58.6 %)	4750 (45.8 %)	<0.01
Chinese/ other Asian	13 (1.1 %)	103 (1.0 %)	0.76
African	9 (0.8 %)	90 (0.9 %)	0.87
Black Caribbean	2 (0.2 %)	36 (0.3 %)	0.43
Mixed race	59 (4.9 %)	647 (6.2 %)	0.07
White British	267 (22.3 %)	3771 (36.4 %)	<0.01
White others	29 (2.4 %)	267 (2.6 %)	0.85
Others	8 (0.7 %)	104 (1.0 %)	0.35
Frequency of diseases, n (%)			
Allergic disease	601 (47.2 %)	4665 (37.3 %)	<0.01
Appendicitis	27 (2.1 %)	24 (0.2 %)	<0.01
Arthritis	1 (0.1 %)	6 (0.0 %)	0.49
Celiac disease	9 (0.7 %)	24 (0.2 %)	<0.01
Constipation	57 (4.5 %)	142 (1.1 %)	<0.01

FD	4 (0.3 %)	3 (0.0 %)	<0.01
GORD	66 (5.2 %)	342 (2.7 %)	<0.01
IBD	2 (0.2 %)	4 (0.0 %)	0.10
IBS	8 (0.6 %)	9 (0.1 %)	<0.01
Migraine	16 (1.3 %)	53 (0.4 %)	<0.01
Mother’s abdominal pain	631 (49.5 %)	3985 (31.8 %)	<0.01
Mother’s allergic disease	609 (47.8 %)	4931 (39.4 %)	<0.01
Mother’s appendicitis	13 (1.0 %)	94 (0.8 %)	0.31
Mother’s arthritis	109 (8.6 %)	495 (4.0 %)	<0.01
Mother’s Celiac disease	14 (1.1 %)	73 (0.6 %)	0.04
Mother’s chronic fatigue syndrome	3 (0.2 %)	25 (0.2 %)	0.74
Mother’s chronic muscle pain	30 (2.4 %)	167 (1.3 %)	0.01
Mother’s constipation	10 (0.8 %)	128 (1.0 %)	0.55
Mother’s depressive disorder, bipolar disorder	309 (24.3 %)	2632 (21.0 %)	0.01
Mother’s FD (%)	33 (2.6 %)	162 (1.3 %)	<0.01
Mother’s GORD	171 (13.4 %)	876 (7.0 %)	<0.01
Mother’s IBD	10 (0.8 %)	84 (0.7 %)	0.59
Mother’s IBS (%)	126 (9.9 %)	877 (7.0 %)	<0.01
Mother’s migraine	285 (22.4 %)	1845 (14.7 %)	<0.01
Mother’s obsessive-compulsive disorder (%)	10 (0.8 %)	61 (0.5 %)	0.15
Mother’s personality disorder	0 (0.0 %)	21 (0.2 %)	0.25
Mother’s schizophrenia	2 (0.2 %)	6 (0.0 %)	0.17

* 311 values are missing in route of birth. ** 2223 values are missing in ethnicity.

SD, standard deviation; FD, functional dyspepsia; GORD, gastro-oesophageal reflux disease; IBD, inflammatory bowel disease; IBS, irritable bowel syndrome

Table 2. Significant results of the Cox proportional hazards model for the diagnosis of children's abdominal pain

Children's factors	**HR**	**95%CI lower**	**95%CI upper**	**p value**
Female (vs. male)	1.1	1.0	1.3	0.03
Ethnicity (vs. White British)				
Bangladeshi	2.6	2.0	3.4	<0.01
Indian	2.1	1.5	2.9	<0.01
Pakistani	1.8	1.6	2.1	<0.01
Chinese/ other Asian	1.9	1.1	3.2	0.03
White other than British	2.0	1.3	2.9	<0.01
Allergic disease	1.1	1.0	1.3	0.04
Appendicitis	5.1	3.4	7.6	<0.01
Celiac disease	3.0	1.5	5.8	<0.01
Constipation	2.5	1.9	3.3	<0.01
FD	8.2	3.0	22.4	<0.01
GORD	1.5	1.1	1.9	<0.01
IBS	5.1	2.3	11.5	<0.01
Migraine	2.8	1.7	4.5	<0.01
Mother's abdominal pain	1.6	1.4	1.8	<0.01
Mother's allergic disease	1.2	1.0	1.3	0.02
Mother's arthritis	1.4	1.2	1.8	<0.01
Mother's GORD	1.4	1.2	1.7	<0.01
Mother's migraine	1.2	1.1	1.4	0.01

FD, functional dyspepsia; GORD, gastro oesophageal reflux disease; IBD, inflammatory bowel disease; IBS, irritable bowel syndrome

Table 3. Significant results of comparison of clinical characteristics among 3 pain clusters

Cluster No	**Cluster 1 (n = 137)**	**Cluster 2 (n = 677)**	**Cluster 3 (n = 340)**	**P value (Cluster 1 vs. 2)**	**P value (Cluster 1 vs. 3)**	**P value (Cluster 2 vs. 3)**
Allergic disease, n (%)	136 (99.3 %)	465 (68.7 %)	0 (0.0 %)	<0.01	<0.01	<0.01
Celiac disease, n (%)	3 (2.2 %)	1 (0.1 %)	2 (0.6 %)	0.05	0.44	0.78
Mother’s abdominal pain, n (%)	0 (0.0 %)	474 (70.0 %)	157 (46.2 %)	<0.01	<0.01	<0.01
Mother’s allergic disease, n (%)	137 (100.0 %)	271 (40.0 %)	81 (23.8 %)	<0.01	<0.01	<0.01
Mother’s arthritis, n (%)	10 (7.3 %)	77 (11.4 %)	14 (4.1 %)	0.53	0.50	<0.01
Mother’s depressive disorder, bipolar disorder, n (%)	28 (20.4 %)	252 (37.2 %)	0 (0.0 %)	<0.01	<0.01	<0.01
Mother’s FD, n (%)	1 (0.7 %)	27 (4.0 %)	4 (1.2 %)	0.21	1.00	0.04
Mother’s GORD, n (%)	14 (10.2 %)	128 (18.9 %)	14 (4.1 %)	0.04	0.05	<0.01
Mother’s IBS, n (%)	12 (8.8 %)	86 (12.7 %)	24 (7.1 %)	0.75	1.00	0.02
Mother’s migraine, n (%)	15 (10.9 %)	230 (34.0 %)	20 (5.9 %)	<0.01	0.24	<0.01

P values are corrected by Bonferroni correction.

FD, functional dyspepsia; GORD, gastro-oesophageal reflux disease; IBD, inflammatory bowel disease; IBS, irritable bowel syndrome

Table 4. Top 10 single nucleotide polymorphisms with significant adjusted p value detected by genome-wide association study between the pain clusters and control

Clusters (vs. control)	SNP	Chr	Minor/major allele	OR	95% CI	P value	Adjusted p value	Annotated gene
Cluster 1	rs8192907	8	A/G	8.38	3.95-17.78	3.12×10^{-8}	0.01	CYP7B1
	rs10872670	6	A/G	2.60	1.85-3.65	3.74×10^{-8}	0.01	AKAP12
	rs10911382	1	A/G	2.97	1.96-4.50	3.09×10^{-7}	0.03	ARPC5
	rs6636070	23	G/A	2.33	1.68-3.25	5.40×10^{-7}	0.03	SOX3
	rs1946326	3	A/G	2.99	1.94-4.60	6.59×10^{-7}	0.03	CLDN16
	rs984564	7	G/A	2.94	1.92-4.50	6.83×10^{-7}	0.03	PPP1R9A
	rs6535115	4	G/A	2.34	1.67-3.28	9.38×10^{-7}	0.03	RP11-780O17.1
	rs191195577	19	G/A	5.68	2.82-11.44	1.19×10^{-6}	0.03	MLL4
	rs1867219	15	C/A	2.49	1.72-3.60	1.36×10^{-6}	0.03	RCCD1
	rs8042680	15	A/C	2.48	1.72-3.60	1.48×10^{-6}	0.03	PRC1
Cluster 2	rs993147	2	G/A	1.83	1.51-2.23	1.22×10^{-9}	<0.01	MIR548AE1
	rs11866841	16	G/A	1.48	1.27-1.72	2.83×10^{-7}	0.02	BANP
	rs131020	22	A/G	0.65	0.55-0.77	3.01×10^{-7}	0.02	FAM19A5

	rs7766140	6	A/G	1.93	1.50-2.47	3.01×10^{-7}	0.02	EYA4
	rs17150239	7	A/G	1.51	1.29-1.78	3.13×10^{-7}	0.02	OSBPL3
	rs4907333	8	G/A	1.48	1.27-1.72	5.07×10^{-7}	0.02	AC138647.1
	rs4843285	16	A/G	1.46	1.26-1.70	6.43×10^{-7}	0.02	BANP
	rs7985529	13	A/G	1.57	1.31-1.87	8.22×10^{-7}	0.02	RP11-309I15.1
	rs11076377	16	A/G	1.44	1.25-1.67	9.59×10^{-7}	0.02	RP11-354I13.2
	rs7191581	16	G/A	0.69	0.59-0.80	1.19×10^{-6}	0.02	BANP
Cluster 3	rs6723106	2	G/A	0.49	0.39-0.63	1.68×10^{-8}	<0.01	WDSUB1
	rs16948563	17	A/G	2.21	1.67-2.91	1.91×10^{-8}	<0.01	DLX3
	rs7586797	2	A/C	1.81	1.46-2.24	6.18×10^{-8}	<0.01	AC097499.1
	rs1820556	2	A/C	1.81	1.46-2.24	6.45×10^{-8}	<0.01	AC097499.1
	rs7603170	2	G/A	0.50	0.39-0.65	1.16×10^{-7}	<0.01	WDSUB1
	rs7591849	2	A/C	0.54	0.43-0.68	1.19×10^{-7}	<0.01	WDSUB1
	rs7914376	10	A/C	1.89	1.49-2.39	1.34×10^{-7}	<0.01	U7
	rs2032729	18	A/G	1.79	1.44-2.23	2.36×10^{-7}	0.01	ZNF24
	rs1116286	2	A/G	1.76	1.42-2.18	2.64×10^{-7}	0.01	AC097499.1
	rs11676419	2	G/A	0.52	0.40-0.66	2.72×10^{-7}	0.01	BAZ2B

OR, odds ratio; CI, confidence interval; Chr, chromosome; BP, base-pair; MAF, Minor allele frequency; SNP, Single Nucleotide Polymorphism

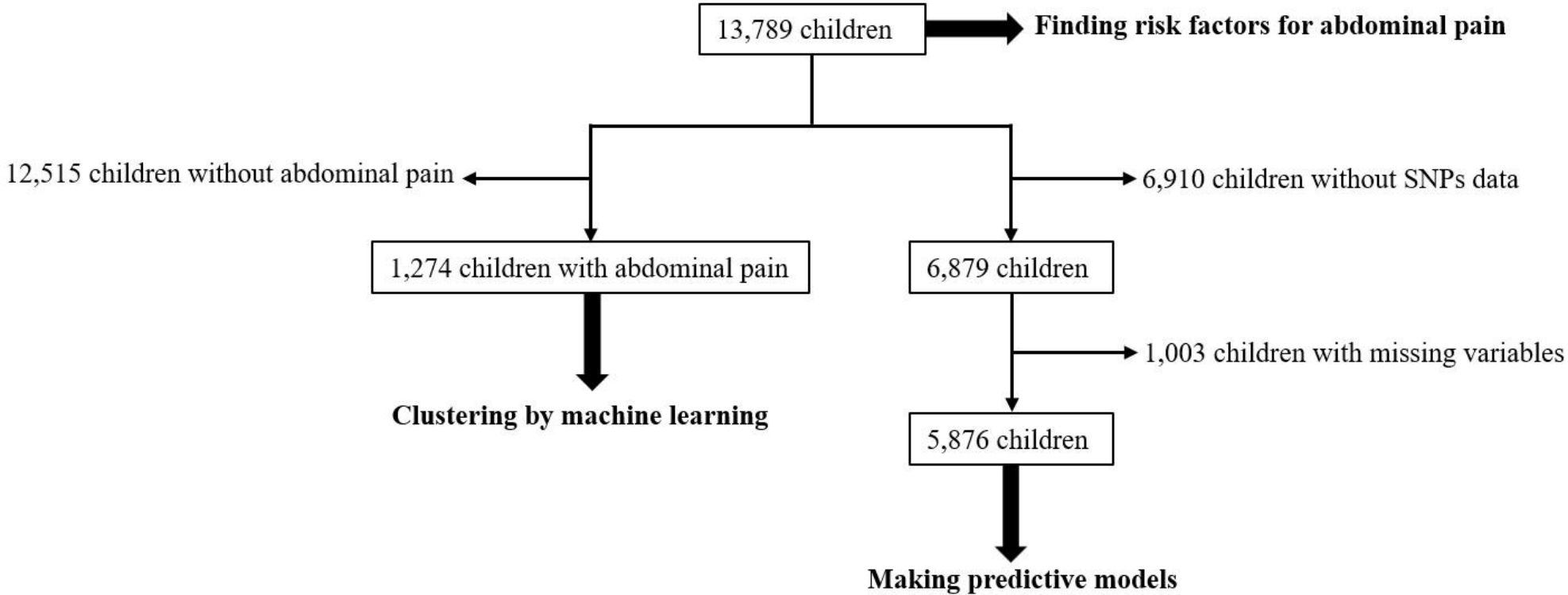
Table 5. Significant results of gene ontology enrichment analysis and pathway enrichment analysis

Cluster	Category	Term	Adjusted p-value	Genes
Cluster 1	Biological Process	N/A	N/A	N/A
	Molecular Function	magnesium ion transmembrane transporter activity (GO:0015095)	0.04	CLDN16
		Adenylate cyclase binding (GO:0008179)	0.04	AKAP12
	Cellular Component	N/A	N/A	N/A
	Pathway enrichment analysis	Tight junction	0.03	ARPC5;CLDN16
		Pathogenic Escherichia coli infection	0.03	ARPC5;CLDN16
		Primary bile acid biosynthesis	0.04	CYP7B1
Cluster 2	Biological Process	negative regulation of extrinsic apoptotic signaling pathway in absence of ligand (GO:2001240)	0.04	EYA4
		negative regulation of signal transduction in absence of ligand (GO:1901099)	0.04	EYA4
		bile acid biosynthetic process (GO:0006699)	0.04	OSBPL3

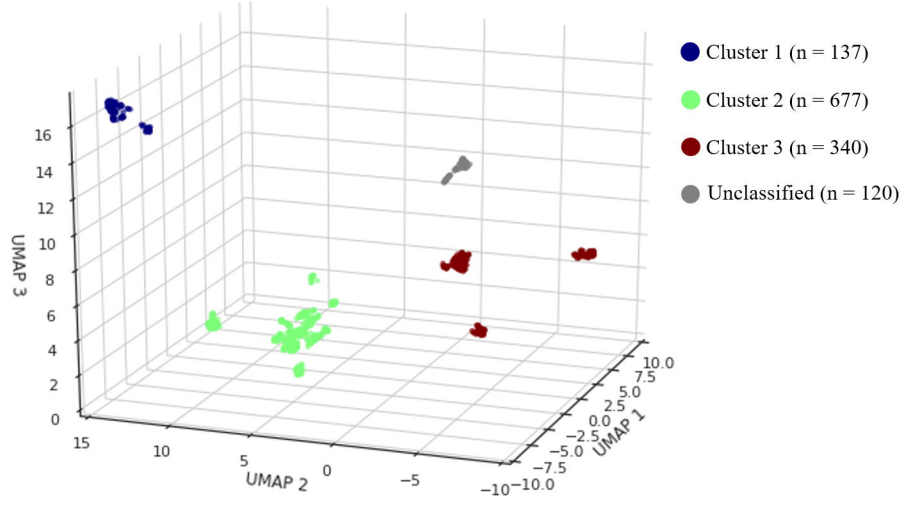
Regulation of extrinsic apoptotic signaling pathway in absence of ligand (GO:2001239)	0.04	EYA4
bile acid metabolic process (GO:0008206)	0.04	OSBPL3
negative regulation of protein catabolic process (GO:0042177)	0.04	BANP
regulation of DNA repair (GO:0006282)	0.04	EYA4
positive regulation of DNA repair (GO:0045739)	0.04	EYA4
negative regulation of catabolic process (GO:0009895)	0.04	BANP
organic hydroxy compound biosynthetic process (GO:1901617)	0.04	OSBPL3
negative regulation of protein metabolic process (GO:0051248)	0.04	BANP
positive regulation of response to DNA damage stimulus (GO:2001022)	0.04	EYA4
monocarboxylic acid biosynthetic process (GO:0072330)	0.04	OSBPL3
positive regulation of DNA metabolic process (GO:0051054)	0.04	EYA4
steroid biosynthetic process	0.04	OSBPL3

		(GO:0006694)		
		negative regulation of extrinsic apoptotic signaling pathway (GO:2001237)	0.05	EYA4
	Molecular Function	cholesterol binding (GO:0015485)	0.03	OSBPL3
		sterol binding (GO:0032934)	0.03	OSBPL3
		protein tyrosine phosphatase activity (GO:0004725)	0.03	EYA4
	Cellular Component	N/A	N/A	N/A
	Pathway enrichment analysis	N/A	N/A	N/A
Cluster 3	Biological Process	N/A	N/A	N/A
	Molecular Function	N/A	N/A	N/A
	Cellular Component	N/A	N/A	N/A
	Pathway enrichment analysis	N/A	N/A	N/A

Adjusted p-values were calculated by Benjamini-Hochberg correction with a threshold for statistical significance $p < 0.05$.
GO, gene ontology; DNA, deoxyribonucleic acid

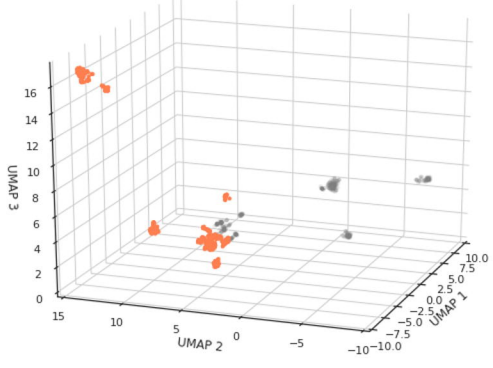

13,789 children
Finding risk factors for abdominal pain
12,515 children without abdominal pain
6,910 children without SNPs data
1,274 children with abdominal pain
6,879 children
1,003 children with missing variables
Clustering by machine learning
5,876 children
Making predictive models

(a) Clustering by UMAP combined with HDBSCAN

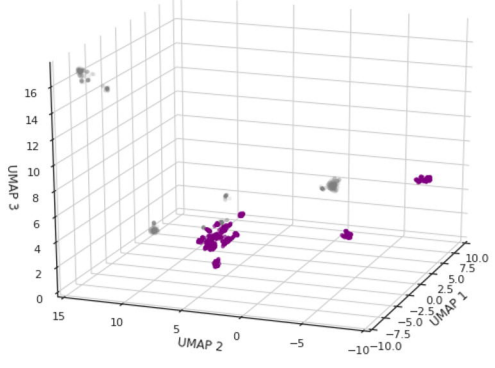


(b) Scattering plots by UMAP labeled by representative features

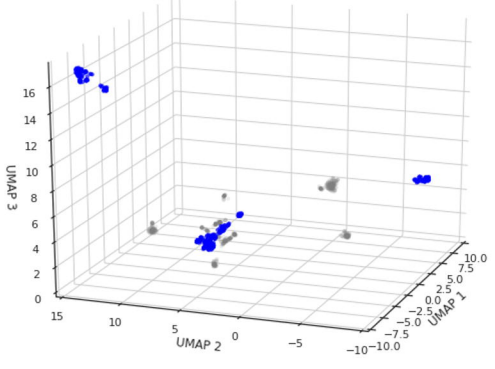
Allergic disease



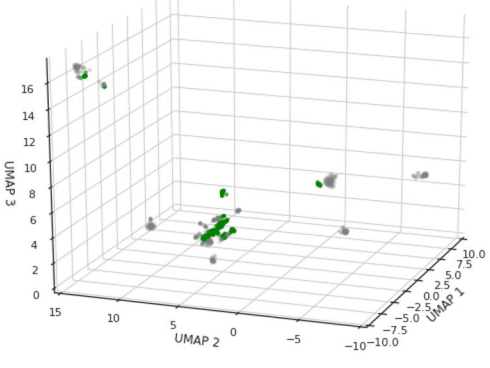
Mother's abdominal pain



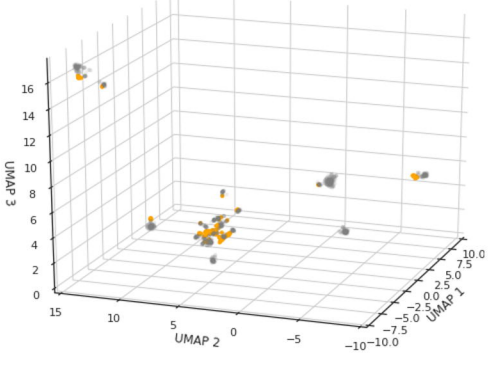
Mother's allergic disease



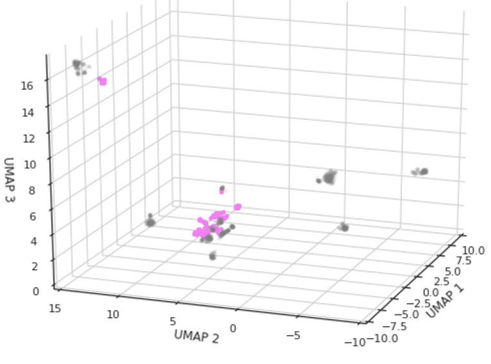
Mother's migraine



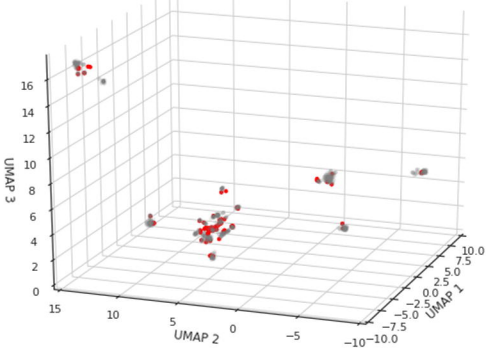
Mother's GORD



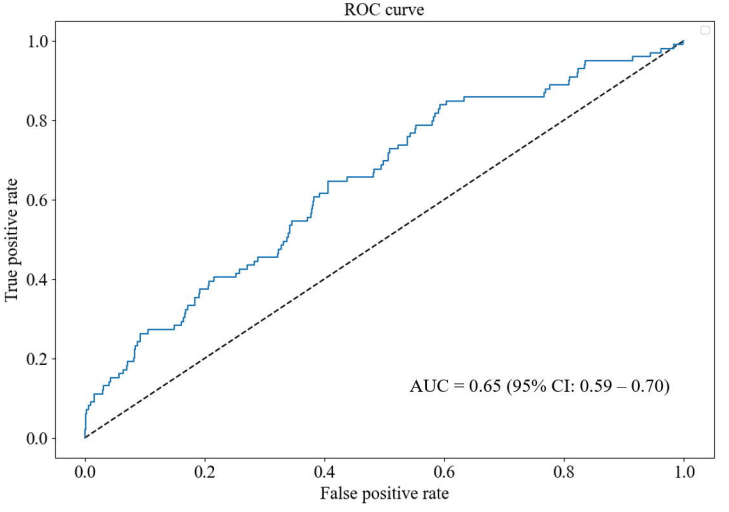
Mother's depressive and bipolar disorder



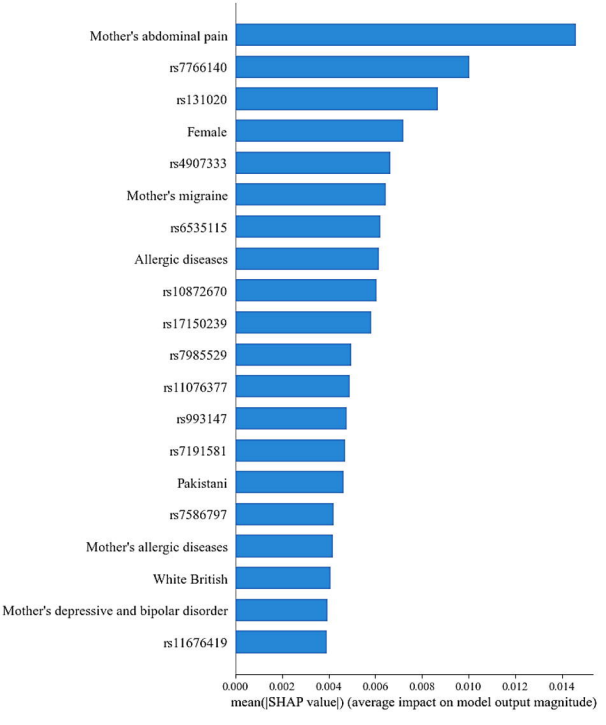
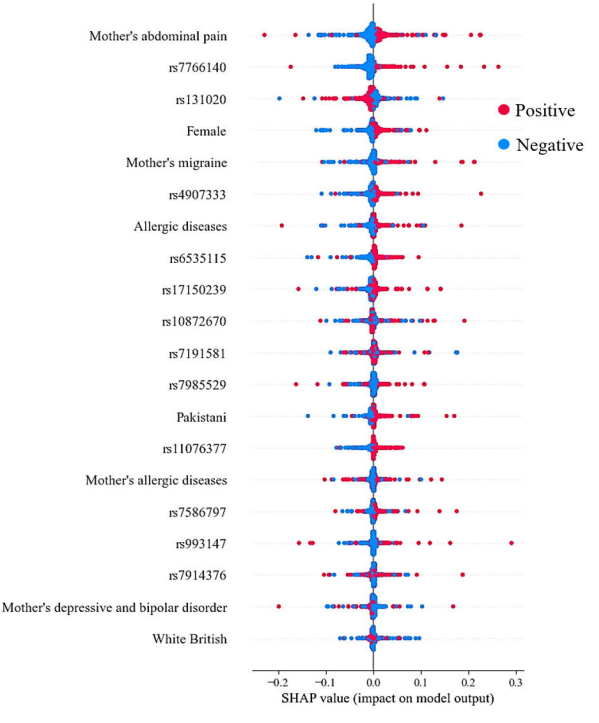
Mother's arthritis



(a) Receiver operating characteristics (ROC) curve and areas under the curve (AUC) of Model 1



(b) The top 20 SHAP feature importance



(c) The effect direction of each feature