

A peer-reviewed version of this preprint was published in PeerJ on 26 July 2018.

[View the peer-reviewed version](https://doi.org/10.7717/peerj.5047) (peerj.com/articles/5047), which is the preferred citable publication unless you specifically need to cite this preprint.

Koci O, Logan M, Svolos V, Russell RK, Gerasimidis K, Ijaz UZ. 2018. An automated identification and analysis of ontological terms in gastrointestinal diseases and nutrition-related literature provides useful insights. PeerJ 6:e5047 <https://doi.org/10.7717/peerj.5047>

An automated identification and analysis of ontological terms in gastrointestinal diseases and nutrition-related literature provides useful insights

Orges Koci¹, Michael Logan², Vaios Svolos¹, Richard K. Russell³, Konstantinos Gerasimidis¹, Umer Zeeshan Ijaz^{Corresp.}²

¹ Human Nutrition, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

² Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, United Kingdom

³ Department of Paediatric Gastroenterology, Royal Hospital for Children, Glasgow, UK, Glasgow, United Kingdom

Corresponding Author: Umer Zeeshan Ijaz

Email address: umer.ijaz@glasgow.ac.uk

With an unprecedented growth in the biomedical literature, keeping up to date with the new developments presents an immense challenge. Publications are often studied in isolation of the established literature, with interpretation being subjective and often introducing human bias. With ontology-driven annotation of biomedical data gaining popularity in recent years and online databases offering metatags with rich textual information, it is now possible to automatically text-mine ontological terms and complement the laborious task of manual management, interpretation, and analysis of the accumulated literature with downstream statistical analysis. In this paper, we have formulated an automated workflow through which we have identified ontological information, including nutrition-related terms in PubMed abstracts (from 1991 until 2016) for two main types of Inflammatory Bowel Diseases: *Crohn's Disease* and *Ulcerative Colitis*; and two other gastrointestinal diseases, namely, *Coeliac Disease* and *Irritable Bowel Syndrome*. Our analysis reveals unique clustering patterns as well as spatial and temporal trends inherent to the considered gastrointestinal diseases in terms of literature that has been accumulated so far. Although automated interpretation cannot replace human judgement, the developed workflow shows promising results and can be a useful tool in systematic literature reviews. The workflow is available at <https://github.com/KociOrges/pytag>.

An automated identification and analysis of ontological terms in gastrointestinal diseases and nutrition-related literature provides useful insights

Orges Koci¹, Michael Logan³, Vaios Svolos¹, Richard K. Russell², Konstantinos Gerasimidis¹, and Umer Zeeshan Ijaz^{3*}

¹Human Nutrition, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow Royal Infirmary, Glasgow, UK

²Department of Paediatric Gastroenterology, Hepatology and Nutrition, Royal Hospital for Children, Glasgow, UK

³School of Engineering, University of Glasgow, Glasgow, UK

*To whom correspondence should be addressed

School of Engineering,

Oakfield Avenue

University of Glasgow

Glasgow

G12 8LT

Umer.Ijaz@glasgow.ac.uk

Tel: +44(0)141-330-6458

Abstract

With an unprecedented growth in the biomedical literature, keeping up to date with the new developments presents an immense challenge. Publications are often studied in isolation of the established literature, with interpretation being subjective and often introducing human bias.

With ontology-driven annotation of biomedical data gaining popularity in recent years and online databases offering metatags with rich textual information, it is now possible to automatically text-mine ontological terms and complement the laborious task of manual management,

interpretation, and analysis of the accumulated literature with downstream statistical analysis. In this paper, we have formulated an automated workflow through which we have identified ontological information, including nutrition-related terms in PubMed abstracts (from 1991 until 2016) for two main types of *Inflammatory Bowel Diseases: Crohn's Disease* and *Ulcerative Colitis*; and two other gastrointestinal diseases, namely, *Coeliac Disease* and *Irritable Bowel Syndrome*. Our analysis reveals unique clustering patterns as well as spatial and temporal trends inherent to the considered gastrointestinal diseases in terms of literature that has been accumulated so far. Although automated interpretation cannot replace human judgement, the developed workflow shows promising results and can be a useful tool in systematic literature reviews. The workflow is available at <https://github.com/KociOrges/pytag>.

Introduction

The volume of biomedical literature in electronic format has grown exponentially over the past few years (Hunter & Cohen, 2006). With the latest count of 27 million in 2017, PubMed search engine can navigate the MEDLINE database of references and abstracts on life and biomedical sciences using key concepts. Lately, ontology-driven annotation of data has become increasingly important, especially in the biomedical domain (Bodenreider & Stevens, 2006; Lambrix et al., 2007). Ontologies describe controlled dictionaries of words on a given theme. By text-mining published abstracts and grouping words used into existing ontologies, it is possible to complement the demanding task of manual management, interpretation, and analysis of the vast amount of available research. Previously (Sinclair et al., 2016), we formulated a pipeline called *seqenv* through which short DNA sequences can be aligned against the NCBI reference databases to extract *Environmental Ontology* (Buttigieg et al., 2013) controlled vocabulary from the metadata (*Isolation Source* field or relevant PubMed abstracts) associated with the matches. Although *prima facie*, one might argue that an abstract is not a full paper, it is still a useful piece of information and with a great number of such short texts (typically in thousands), *seqenv* did indeed show potential in environmental source tracking of DNA sequences. Using the same principle that we applied to sequencing data, in this paper, we developed a new workflow that automatically annotates PubMed abstracts with rich ontological terms. This can be applied to any disease conditions, as well as allowing the user to perform the same search longitudinally, to

highlight changes in a particular area. Downstream data analysis employing ecological statistics is then performed to allow the investigator to interrogate patterns in the context of ontological terms and identify differences between chosen disease groups as well as secular developments within each of these.

This methodology is useful because one not only gets a historical perspective by exploring trends of how the research in a specific topic evolves over a period of time but can also use this information to predict where a particular literature theme is heading. Such an approach can be helpful for systematic reviews as it benefits from the ability to inspect very large number of studies and rapidly annotate thousands of articles with rich metadata in a time-efficient manner of few minutes. It can also reduce the amount of information to manageable sets from which it is easier to infer patterns and trends. In addition, statistical analysis of metadata from multiple ontologies can capture additional details of the content of a research paper and reveal patterns about topics differentiating between test and control groups, something not possible with a traditional, manual approach applied in systematic reviews.

In this paper, we propose a workflow to annotate journal abstracts from nutrition-related literature relevant to two main types of Inflammatory Bowel Diseases (IBDs), namely, *Crohn's Disease* (CD) and *Ulcerative Colitis* (UC); and two other Gastrointestinal (GI) conditions, *Coeliac Disease* (CCD), and *Irritable Bowel Syndrome* (IBS) where it was assumed *a priori* these will stand out in terms of nutrition-related terms from the former. We were particularly interested in the subset of papers that covered aspects related to nutrition, using one or more of the following search keywords: *Diet*, *Food*, and *Nutrition*. We hypothesised that: a) distinct clustering will be observed in nutrition-related terms between the IBD and non-IBD groups; b) there will be a minimal overlap and close proximity of closely associated conditions on an ordination diagram (beta diversity measure) to suggest that specific ontological terms (e.g. underlying aetiology and dietary factors) are differentiating or converging to similar set of principles; c) we will be able to pick up nutrition terms that have gained/lost interest in the disease groups ("V" or "inverted-V" shape curves over time); and d) pinpoint exact location in time when underlying research in terms of nutrition has shifted from exploration (high variability in terms) to exploitation (convergence to certain terms).

Materials and Methods

Search strategy for GI diseases and nutrition-related literature

The abstracts used for analysis were retrieved from PubMed database using the list of keywords described in Figure 1, using a time frame from 1991 until 2016 (searches were performed in July 2017). Composite keywords were constructed through Boolean logic, with four by three possibilities (4 disease groups x 3 nutritional keywords, yielding twelve possible combinations). The returned abstracts were then grouped together in pairs of years, extracted and stored in external files, using the “Citation Manager” function in MEDLINE (tagged) format. The complete search for all the possible combinations from 1991 to 2016, yielded a total number of 24,559 PubMed abstracts. These were then imported into EndNote® X7 citation management software to export them in BibTeX format (input format for our software), where every abstract was described by a number of records including the PubMed ID i.e., a unique identifier used in PubMed and assigned to each article record when it enters the PubMed system. In total, 156 BibTeX files were generated for all the possible combinations of composite keywords and pairs of years (i.e., twelve possibilities in a 26-year timeline).

Annotation Process

The BibTeX files were then processed with our novel pyTag workflow. Using pyTag allows the relevant abstract, from each PubMed ID to be extracted from the NCBI database and collated together for a given group, e.g. the one describing the literature for Crohn’s Disease. Next, these abstracts were annotated using a custom named entity recognition (NER) system, i.e. a method for the automatic identification of ontological terms mentioned in texts, called EXTRACT (2.0, Pafilis, Bērziņš & Jensen, 2017). The system supports multiple ontologies (a controlled dictionary of words on a given theme) and can recover mentions for *Organisms* (NCBI Taxonomy, Federhen, 2011), *Environments* (Environment Ontology, Buttigieg et al., 2016), *Diseases and phenotypes* (Disease Ontology, Kibbe et al., 2014; Mammalian Phenotype Ontology, Smith & Eppig, 2012), *Tissues and cell lines* (BRENDA Tissue Ontology, Placzek et

al., 2016), *Biological processes, molecular functions, and cellular components* (The Gene Ontology Consortium, 2015), *Genes/Proteins* (STRING, Szklarczyk et al., 2016; RAIN, Junge et al., 2017) and *Small molecule compounds* (STITCH, Szklarczyk et al., 2015) in a given piece of text. After the annotation of the total number of abstracts, the resulting frequency of the identified terms was converted to a two-dimensional abundance table, with enough replicates per group to ensure that ecological statistics including alpha and beta diversities could be calculated as well as differential analysis could be performed. This is summarized in Figure 2. For the annotation of the literature, all the ontologies supported by the system were employed. Out of 24,559 abstracts, 21,035 of them were annotated, i.e. at least one term was found in their content (for terms appearing more than once in an abstract only one occurrence was considered). From the identified terms, those with low or rare frequencies were removed (< 5 total hits across all searches). From the remaining 2,399, 445 terms relevant only in the context of nutrition were selected using a manually developed nutrition-only ontology library and these were considered for statistical analysis. Therefore, in this study where we use the word “terms” it is implicitly assumed that they are relevant to nutrition only.

Statistical Analysis

Statistical analysis was performed in R software. To account for the variation of the number of publications over time, the counts of each term, found in a search for a pair of years for a specific disease condition, were adjusted with respect to the number of the papers published in literature for this condition and annotated from the workflow for this specific year (document-based normalisation). To explore the significance of the variability of ontological terms between the disease conditions, the Vegan package (Oksanen et al., 2017) was used, particularly, the function *adonis* for PERMANOVA (ANOVA for distance matrices). Clustering between the disease groups, how dissimilar the terms for a given search (e.g., year or condition) are from each other and temporal changes in literature were assessed using the reduced-order representation of the datasets using the non-metric multidimensional scaling (NMDS), which reduces the multivariate dataset to two or more dimensions (similar to PCA) based on dissimilarity (Bray-Curtis distance) between the terms for a given search. The Local Contributions to Beta Diversity (LCBD) was also used with a Hellinger transformation (Legendre & De Cáceres, 2013), where the overall beta

diversity is divided into individual contributions from samples to identify outliers. The smaller the LCBD value is, the closer the sample is to the group average. To identify ontology-based terms that were significantly different between the conditions, *Kruskal-Wallis* test (Kruskal & Wallis, 1952) was used. The *Benjamini-Hochberg* correction was used on the returned p-values to correct for multiple testing and *Dunn's* test as a post-hoc procedure for pair-wise comparisons, where appropriate.

Results

Ontological terms clustered IBD separately from non-IBD conditions with temporal changes observed in the literature of each disease group

When the composition of the ontological terms for the disease conditions was assessed using NMDS plots, findings demonstrated an evident clustering of IBD related ontological terms distinct from non-IBD (Fig. 3A). The clusters for CCD and IBS stood well apart from those of CD and UC. CD and UC showed a degree of overlap, suggesting a degree of similarity in the ontological terms between these two conditions. Temporal variability was also noticeable from the NMDS plot (Fig. 3B). The beta diversity analysis revealed that the nutrition-related literature for each disease group has shifted over time. For all groups, the between-year variability was higher in the earlier dates, but gradually decreased, as we moved forward in time. This was clearer in the case of CD, UC and IBS. It could be seen that the proximity between CD and UC was increasing more for the later years and that the two IBD groups were further converging to a similar set of ontological terms.

The convergence between the groups was also obvious when LCBD (Legendre & De Cáceres, 2013) analysis was applied. The findings, in this case, showed a decreasing trend of the LCBD values over the years for each disease group, more noticeable for the case of CD, UC and IBS (Fig. 4). This indicated that the relative contribution of each sample (search for a pair of years) in every group was shifting towards the mean value (multivariate centroid) of the sample space when approaching more recent dates, suggesting their gradual convergence in recent years. This

pattern indicated a relative consensus on a particular nutrition research theme for these disease conditions.

Most frequent topics and conserved patterns in the literature of the disease conditions

PERMANOVA (distances between groups) suggested that most of the variability was explained significantly by the different disease conditions ($R^2 = 27\%$, $p = 0.001$). To further explore this and inspect for terms that stratify the groups, we first looked at the twenty most frequent terms in the literature of each condition for the entire time frame. Findings showed that CD and UC, shared more than a half (65%) of their most common topics, and terms such as *growth* (Freq. CD = 3.90; UC = 3.00) and *fatty acids* (Freq. CD = 2.08, UC = 2.78) were listed as the top two most frequent in the literature of the IBD groups (Fig. 5). In a similar way, *wheat* (Freq. = 6.16) and *gliadin* (Freq. = 5.12) were unsurprisingly some of the most prevalent in the literature of CDD research (Fig. 5). Likewise, the ontological terms *fibre* (Freq. = 4.57) and *lactose* (Freq. = 3.10) were found very common in IBS (Fig. 5).

Moreover, differential analysis performed over separate time intervals (see Table 1) showed that the above findings were fairly conserved between the groups over the years (Fig. 6). This can suggest a continuous scientific interest for these topics in the research of each disease. In addition, results showed multiple terms becoming significant between the disease conditions for each time interval ($P_{adj} < 0.05$; see Table 1 and Tables S1A-D). Specifically, in CCD, terms for *gliadin*, *wheat*, *rye*, *barley* and *oats* were found to be stably frequent between 1991 and 2016, and clearly more common compared to the other groups ($CCD > \text{other diseases}$; Fig. 6 and Tables S1A-D). In a similar way, a considerable presence of *fibre* and *lactose* was observed in IBS throughout the years with findings also indicating a decrease in the frequency of both terms for the more recent dates (Fig. 6).

In the case of the IBD groups, terms such as *omega-3 fatty acids* and *n-6 fatty acids* were evidently more frequent compared to IBS and CCD where they were less common ($CD \text{ and } UC > CCD \text{ and } IBS$; Fig. 6 and Tables S1B-D). For *omega-3 fatty acids*, the pattern was relatively stable over time (between 1991 and 2016) with some slight decrease for both CD and UC

between 2011 and 2016 (Fig. 6). In a similar way, *n-6 fatty acids* were very common in CD and UC between 1999 and 2016 (Tables S1B-D). *Growth* term was also observed to be significantly different between the disease groups (Fig. 6). In CD, the same term had the highest prevalence with UC and CCD following respectively, appearing the least in IBS (CD > UC > CCD > IBS; Fig. 6). However, only during 1991-1998, this term appeared in CCD almost in similar levels to CD and UC literature.

Ontological terms showing temporal changes in the literature of the disease groups

Analysis of variance using the *adonis* function showed that also temporal variability (expressed as in pairs of years) explained up to 19% of the changes in the use of ontological terms ($R^2 = 19.0\%$, $p = 0.001$). To investigate this further, differential analysis was performed on each term (see Table 1). Findings showed a number of terms differentiating over time in the literature of the disease conditions ($P_{adj} < 0.05$; see Table 1 and Tables S2A-D).

More specifically, results revealed a considerable increase in the frequency of *obesity* term for all disease conditions (P_{adj} CCD = 0.025, CD = 0.01083, IBS = 0.00691, UC = 0.01108; Fig. 7A). This was more evident after years 2008-09, for each disease group. *Obesity* was higher in IBS between 2009-10 and 2015-16 compared to the other groups, with CCD being next, and CD and UC following respectively. Similarly, *wheat allergy* was found becoming more common between the disease conditions over the years (P_{adj} CCD = 0.01712, CD = 0.03986, IBS = 0.00381, UC = 0.04184; Fig. 7B). This term was noticed more frequent for CCD and IBS in the more recent dates (2011-12 and thereafter). CCD seemed to be the group where *wheat allergy* was increasing the most with IBS being next. In the case of CD and UC, the same term was found to be equally prevalent between 2015-16 for both groups, but clearly in a lower frequency when compared to the non-IBD types.

The frequency of several terms was also found to change temporally in relation to CD and UC (Fig. 7C). This was the case for *butyrate* (P_{adj} CCD = 0.025, UC = 0.01108) and *curcumin* (P_{adj} CCD = 0.01549, UC = 0.04184). *Butyrate* showed an increasing trend in the literature, most prominently in UC, with a peak frequency noticed in 2001-02, and becoming considerably less

common onwards (Fig. 7C). The same term was notably less common in CD compared to UC, where it became frequent between 1999-00 and 2001-02 and it was found in similar levels to UC in 2015-16 (Fig. 7C). In addition, a partially transient prevalence over time was seen for the term *short-chain fatty acids (SCFAs)*. *SCFAs* (Padj CCD = 0.01404, UC = 0.01763) were noticed to be more frequent for both groups between 1993-94 and 2003-04 and decreasing rapidly onwards, particularly in the case of UC (Fig. 7C). Moreover, the term *vitamin D* (Padj CCD = 0.01242, UC = 0.04184) was found more common in CD compared to UC and becoming frequent over the years for both groups showing a notable increase between 2009-10 and 2013-14 (Fig. 7C). However, after these dates, a slight decrease could be observed in both cases for the years 2015-16.

Discussion

In this study, we collated and assessed nutrition-related ontological terms from the literature of IBD and two other gastrointestinal conditions. We inspected how certain nutrition terms differentiated between the groups and evolved in the scientific literature over the last 26 years. Results showed discriminating differences between IBD and non-IBD types and secular patterns in the literature of each disease separately. It was demonstrated that the terms related to the IBD types clustered distinctly from those of the non-IBDs. It was shown that the literature of each group was shifting over time and that it was gradually converging for the recent dates in the timeline. This was more evident for the case of CD and UC, but also noticeable for the other groups as well. This suggests that research topics are similar in the recent years for these diseases.

The prevalence of several terms that stratify the disease conditions in a conserved manner over time was also illustrated. More specifically, it was clearly noticed that terms describing gluten-related proteins and containing food, such as *gliadin*, *wheat*, *rye* and *barley* were found in high frequencies in the literature of CCD. This was an expected outcome for CCD (McGough & Cummings, 2005) and suggests that our workflow is specific. Similarly, *fibre* was found to be considerably prevalent for IBS compared to the other groups. This observation aligns with studies suggesting that alteration of certain dietary *fibre* intake can be beneficial for this

condition (El-Salhy et al., 2012), and a low FODMAP diet is now recognised as a successful management strategy for functional bowel disorders like IBS (Staudacher et al., 2011; Halmos et al., 2014). In the case of the IBD, terms such as *omega-3 fatty acids* and *n-6 fatty acids* were very common compared to IBS and CCD where their frequency was very low. This finding aligns to studies exploring the role of *omega-3* and *n-6 fatty acids* in the regulation of inflammation and as treatment modalities in IBD (Cabr , Ma osa & Gassull, 2012; Patterson et al., 2012; Barbalho et al., 2016;), although their clinical efficacy is now less clear. In addition, the frequency of *growth* term appeared more prominently in the IBD groups compared to the other conditions, and more evidently in the case of CD, where height deficits are more often compared with UC or IBS where delayed *growth* and short stature are less common (Gerasimidis, McGrogan & Edwards, 2011; Sigall-Boneh et al., 2017; Mason et al., 2017).

Patterns from temporal analysis revealed that *obesity* was steadily increasing in all groups and becoming very common in literature. This finding is in agreement with recent evidence from studies showing a growing prevalence of *obesity* in IBD patients (Flores et al., 2015) and mechanistic studies trying to unravel the role of adipose tissue in the inflammatory response (Wozniak et al., 2008; Bertin, Desreumaux & Dubuquoy, 2010). In the past, while malnutrition and inadequate nutrition in CD and UC patients were studied as the most common extra-intestinal complications in IBD, research seems to shift to studies looking at *overnutrition* and *obesity*.

On the contrary, a transient focus was demonstrated for *short-chain fatty acids* and particularly *butyrate*, in both UD and CD. *SCFAs* are well known and characterised bacterial metabolites produced from the fermentation of undigested fibre in the colon. The level of *SCFAs* content in faecal samples has been shown to be related to the pathogenesis of some gastrointestinal conditions, including IBD (Venter, Vorster & Cummings, 1990). Among *SCFAs*, *butyrate* is the most extensively studied, and several clinical studies document beneficial effects of *butyrate* but also issues with its production and colonic utilization in IBD (Scheppach et al., 1992; Steinhart et al., 1996). However, the frequencies of both these terms were found to become considerably lower, especially in the case of UC, for the more recent years reflecting a loss of interest in these topics in IBD research. This trend may represent the evolution of microbiome research in IBD

from the role certain metabolites to the broader role of the microbiome and its broad metabolites, particularly now that OMICS technologies and computational power are more accessible. An interesting trend was seen for *vitamin D*. Despite the steady increase been observed for this term over time, a decrease of published interest has been noticed recently, in both IBD groups. This observation is likely to indicate an increase in the role of *vitamin D* in IBD pathogenesis, considering particularly the high prevalence in this population, which has recently declined in the absence of consistent evidence implicating this vitamin as an environmental risk factor for autoimmune diseases like CD (Narula & Marshall, 2012). The decrease found in the recent years hence may suggest that less clinical attention is now given to the role of *vitamin D* in IBD or that this certain research theme has been exhaustively studied.

Conclusions

We have presented a rapid, automated workflow for the systematic annotation of scientific literature with rich metadata employing a broad range of domain ontologies. We have applied this tool for the identification and analyses of ontological terms in certain gastrointestinal diseases and nutrition-related literature. Although automated interpretation cannot completely replace human judgement, it can save significant time to process very large amounts of literature, free from reviewer's bias, and can reduce this information to a far more comprehensive and manageable set of deducable patterns from which it is easier to draw conclusions. Application of summary statistics, regularly used in environmental microbiology, allow description of differences between multiple conditions and patterns over time within a certain condition. The current workflow is applicable to any type of literature and can perform equally for any kind of published data accessed from PubMed database. However, the manually developed nutrition-only ontology library used in this study highlights the need to develop theme specific ontology libraries that can make the workflow more effective and more efficient.

Competing Interests

The authors declare no competing interest for this study.

Data availability

The code for pyTag workflow and the associated data are available at:
<https://github.com/KociOrges/pytag>.

Author's Contributions

UZI and KG designed the study; UZI, KG, and RR directed this study; OK wrote the software and carried out the statistical analysis; OK and UZI wrote the manuscript; KG critically interpreted findings; VS, ML, KG, and RR provided feedback on the manuscript and clinical relevance of this work; All authors read, commented on, and approved the paper.

Funding

UZI is funded by NERC IRF NE/L011956/1. OK is supported by Nestle Industrial PhD Partnership with the University of Glasgow.

References

- Barbalho SM, Goulart R de A, Quesada K, Bechara MD, de Carvalho A de CA. 2016. Inflammatory bowel disease: can omega-3 fatty acids really help? *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology* **29**:37-43.
- Bertin B, Desreumaux P, Dubuquoy L. 2010. Obesity, visceral fat and Crohn's disease. *Current Opinion in Clinical Nutrition and Metabolic Care* **13**:574-580.
- Bodenreider O, Stevens R. 2006. Bio-ontologies: current trends and future directions. *Brief Bioinformatics* **7**:256-274.

368

369 Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE. 2013. The environment ontology:
370 contextualising biological and biomedical entities. *Journal of Biomedical Semantics* **4** Article 43.

371

372 Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. 2016. The
373 environment ontology in 2016: bridging domains with increased scope, semantic density, and
374 interoperation. *Journal of Biomedical Semantics* **7**

375

376 Cabré E, Mañosa M, Gassull MA. 2012. Omega-3 fatty acids and inflammatory bowel diseases –
377 a systematic review. *British Journal of Nutrition* **107**:S240-S252.

378

379 El-Salhy M, Gundersen D, Hatlebakk JG, Hausken T. 2012. *Irritable Bowel Syndrome:
380 Diagnosis, Pathogenesis and Treatment Options*. New York: Nova Science Publishers.

381

382 Federhen S. 2011. The NCBI Taxonomy database. *Nucleic Acids Research* **40**:D136-D143.

383

384 Flores A, Burstein E, Cipher DJ, Feagins LA. 2015. Obesity in Inflammatory Bowel Disease: A
385 Marker of Less Severe Disease. *Digestive Diseases and Sciences* **60**:2436-2445.

386

387 Gerasimidis K, McGrogan P, Edwards CA. 2011. The aetiology and impact of malnutrition in
388 paediatric inflammatory bowel disease. *Journal of Human Nutrition and Dietetics* **24**:313-326.

389

390 Halmos EP, Power VA, Shepherd SJ, Gibson PR, Muir JG. 2014. A Diet Low in FODMAPs
391 Reduces Symptoms of Irritable Bowel Syndrome. *Gastroenterology* **146**:67-75.e5.

392

393 Hunter L, Cohen KB. 2006. Biomedical Language Processing: What's Beyond
394 PubMed?. *Molecular Cell* **21**:589-594.

395

396 Junge A, Refsgaard JC, Garde C, Pan X, Santos A, Alkan F, Anthon C, von Mering C, Workman
397 CT, Jensen LJ, Gorodkin J. 2017. RAIN: RNA–protein Association and Interaction Networks.
398 Database: *The Journal of Biological Databases and Curation* **2017**:baw167.

399

400 Kibbe WA, Arze C, Felix V, Mitranka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J,
401 Vasant D, Parkinson H, Schriml LM. 2014. Disease Ontology 2015 update: an expanded and
402 updated database of human diseases for linking biomedical knowledge through disease
403 data. *Nucleic Acids Research* **43**:D1071-D1078.

404

405 Kruskal WH, Wallis WA. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of*
406 *the American Statistical Association* **47**:583-621.

407

408 Lambrix P, Tan H, Jakonienė V, Strömbäck L. 2007. Biological Ontologies. In: Baker CJO,
409 Cheung KH, ed. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. New
410 York: Springer, 85-89.

411

412 Legendre P, De Cáceres M. 2013. Beta diversity as the variance of community data: dissimilarity
413 coefficients and partitioning. *Ecology Letters* **16**:951-963.

414

415 Mason A, Gerasimidis K, Iljuhhina J, Laird S, Munro J, Gaya DR, Russell RK, Ahmed SF. 2017.
416 Long-Term Skeletal Disproportion in Childhood-Onset Crohn's Disease. *Hormone Research in*
417 *Paediatrics* **89**:132-135.

418

419 McGough N, Cummings JH. 2005. Coeliac disease: a diverse clinical syndrome caused by
420 intolerance of wheat, barley and rye. *Proceedings of the Nutrition Society* **64**:434-450.

421

422 Narula N, Marshall JK. 2012. Management of inflammatory bowel disease with vitamin D:
423 Beyond bone health. *Journal of Crohn's and Colitis* **6**:397-404.

424

425 Oksanen, J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlin D, Minchlin PR, O'Hara
426 RB, Simpson GL, Solymos P. 2017. Package 'vegan'. *Community Ecology Package, version 2.4-*
427 *3*.

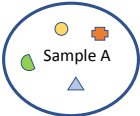
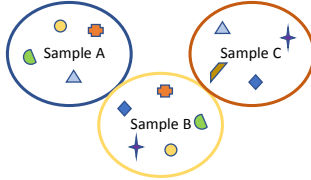
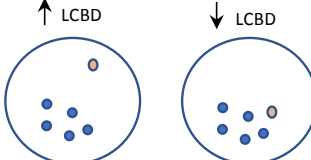
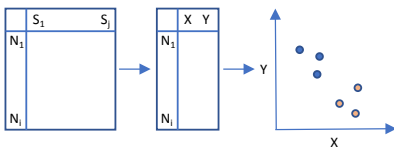
428

- 429 Pafilis E, Bērziņš R, Jensen LJ. 2017. EXTRACT 2.0: text-mining-assisted interactive annotation
430 of biomedical named entities and ontology terms. biorxiv.org preprint
431 <https://doi.org/10.1101/111088>.
432
- 433 Patterson E, Wall R, Fitzgerald GF, Ross RP, Stanton C. 2012. Health Implications of High
434 Dietary Omega-6 Polyunsaturated Fatty Acids. *Journal of Nutrition and Metabolism*
435 **2012**:539426.
436
- 437 Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D. 2016.
438 BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research*
439 **45**:D380-D388.
440
- 441 Scheppach W, Sommer H, Kirchner T, Paganelli GM, Bartram P, Christl S, Richter F, Dusel G,
442 Kasper H. 1992. Effect of butyrate enemas on the colonic mucosa in distal ulcerative
443 colitis. *Gastroenterology* **103**:51-56.
444
- 445 Sigall-Boneh R, Levine A, Lomer M, Wierdsma N, Allan P, Fiorino G, Gatti S, Jonkers D,
446 Kierkuś J, Katsanos KH, Melgar S, Yuksel ES, Whelan K, Wine E, Gerasimidis K. 2017.
447 Research Gaps in Diet and Nutrition in Inflammatory Bowel Disease. A Topical Review by D-
448 ECCO Working Group [Dietitians of ECCO]. *Journal of Crohn's and Colitis* **11**:1407-1419.
449
- 450 Sinclair L, Ijaz UZ, Jensen LJ, Coolen MJL, Gubry-Rangin C, Chroňáková A, Oulas A, Pavloudi
451 C, Schnetzer J, Weimann A, Ijaz A, Eiler A, Quince C, Pafilis E. 2016. Seqenv: linking
452 sequences to environments through text mining. *PeerJ Preprints* **4**:e2317v1.
453
- 454 Smith C, Eppig J. 2012. The Mammalian Phenotype Ontology as a unifying standard for
455 experimental and high-throughput phenotyping data. *Mammalian Genome* **23**:653-668.
456
- 457 Staudacher HM, Whelan K, Irving PM, Lomer MC. 2011. Comparison of symptom response
458 following advice for a diet low in fermentable carbohydrates (FODMAPs) versus standard

- dietary advice in patients with irritable bowel syndrome. *Journal of Human Nutrition and Dietetics* **24**:487-495.
- Steinhart AH, Hiruki T, Brzezinski A, Baker JP. 1996. Treatment of left-sided ulcerative colitis with butyrate enemas: a controlled trial. *Alimentary Pharmacology and Therapeutics* **10**:729-736.
- Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45**:D362-D368.
- Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. 2015. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research* **44**:D380-D384.
- The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Research*. **43**:D1049-D1056.
- Venter CS, Vorster HH, Cummings JH. 1990. Effects of dietary propionate on carbohydrate and lipid metabolism in healthy volunteers. *The American Journal of Gastroenterology* **85**:549-553.
- Wozniak SE, Gee LL, Wachtel MS, Frezza EE. 2008. Adipose Tissue: The New Endocrine Organ? A Review Article. *Digestive Diseases and Sciences* **54**:1847-1856.

Glossary

Terminology	Description	Usefulness	References
Alpha Diversity	Reflects the within-sample diversity.	Inspect how many different individuals e.g., microbial species could be detected	-

		in one sample.	
Beta Diversity	<p>Reflects the between-sample diversity.</p> 	Inspect dissimilarities (distance and/or clustering) between samples.	-
Kruskal-Wallis Test	Test whether the medians of two or more groups are equal.	Determine if there are statistically significant differences between multiple groups (two or more).	R's <i>stats</i> : <code>kruskal.test()</code>
Local Contributions to Beta Diversity (LCBD)	<p>The overall beta diversity is divided into individual contributions from samples. Smaller the LCBD value, the more closer the sample is to the group average.</p> 	Inspect how far or close are the individual contributions from samples to the group average.	-
Non-metric Multidimensional Scaling (NMDS)	<p>Ordination technique where data from multiple dimensions (e.g., from multiple communities, sites, etc.) are simplified into just a few and represented as points in a 2D space (similar to PCA).</p> <p>N: sites, S: species</p> 	Inspect beta diversity of a multivariate dataset in a 2D space.	R's <i>vegan</i> : <code>metaMDS()</code>
Ontology	A formal specifications of a list of terms that were arranged in a hierarchical structure with a unique ID assigned to a term including its' synonyms. A term	Create a consensual controlled vocabulary of terms.	-

	itself can be a part of multiple hierarchies.		
Permutational multivariate analysis of variance (PERMANOVA)	Compare groups of objects and test if there are differences in the position and/or spread, in a multivariate space, of the compared groups attributes.	Measure effect size and significance on beta diversity for a grouping variable.	R's <i>vegan</i> : adonis()

486

487

Table 1. Significance analyses performed on the identified ontological terms. Temporally changing terms were explored for each disease group individually (Subset size). Ontological terms becoming significant between the groups were also explored using differential analysis in separate time intervals. An adjusted p-value (P_{adj}) < 0.05 was considered significant in each test. Percentage indicates the number of terms found significant over the size of the subset used for significance testing. n = total number of nutrition-related terms in the initial composite frequency table.

List of figures

Figure 1. Schematic of the keywords searched in PubMed search engine for the gastrointestinal diseases and nutrition-related literature. Twelve possibilities (4 X disease groups by 3 X nutritional categories) were searched in a 26-year timeline. The returned abstracts were grouped together in pairs of years and collated for a given group.

Figure 2. Schematic of the workflow for the automated identification and analyses of ontological terms in literature data. The abstracts returned from a keyword search in PubMed database are extracted and then processed with the pyTag workflow, where all the ontological terms are listed and annotated. After the annotation, a frequency table of the identified terms is generated and next subjected to statistical analysis.

Figure 3. Non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance demonstrating clustering of IBD and non-IBD groups in the 26-year timeline. Points indicate searches in pairs of years. A) Ellipses describe 95% CI of standard deviation for a given group. B) Dashed arrow represents transitions in the timeline. The size of the points corresponds to the date they describe where smaller size indicates earlier years and larger one more recent dates. SD=Starting date (1991-1992), ED=Ending Date (2015-2016), other=intermediate dates.

Figure 4. The relative contributions to beta diversity (LCBD) per disease condition. LCBD analysis demonstrating temporal variations in the literature of each disease group (distances from group average). Loess curve with shaded 95% CI illustrates the trends for each disease condition.

Figure 5. Top 20 most frequent ontological terms in the literature of each disease condition for the entire time frame (1991-2016).

Figure 6. Ontological terms whose frequency differentiated between the disease groups, over separate subsets of time intervals. Box plots indicate the median, lower and upper quartiles of the document-based normalised frequency obtained for a specific term from the searches performed over the dates of a time interval, across the nutritional categories: Nutrition, Food and Diet, for a single group. Filled circles represent outliers. Dunn's comparison with asterisks indicating significant differences $*=p<0.05$, $**=p<0.01$ and $***=p<0.001$.

Figure 7. Trends of ontological terms whose frequency differentiated temporally in the literature of the gastrointestinal conditions. Plots A) and B) describe the prevalence over time of *obesity* and *wheat allergy* respectively, in all disease groups, and plot C) describes the prevalence of terms found to differentiate over time in relation with CD and UC. Points indicate the mean document-based normalised frequency obtained for a specific term from a search conducted for a pair of years across the nutritional categories: Nutrition, Food and Diet, for a single disease group.

Supplemental Information

Tables_S1.xlsx. Differential expression analysis of nutrition-related terms between disease conditions. Four tables, Table_S1A (1991-1998), Table_S1B (1999-2004), Table_S1C (2005-2010), and Table_S1D (2011-2016) for differential expression analysis of nutrition-related terms between diseases using Kruskal-Wallis test. Only those terms are shown where the adjusted p-value (P_{adj}) < 0.05 . Mean expression indicates the mean document-based normalised frequency obtained for a specific term for each disease group. A post hoc pairwise Dunn's comparison indicating significant differences between the groups is shown on the right half.

Tables_S2.xlsx. Differential expression analysis of nutrition-related terms between years. Four tables, Table_S2A (CCD), Table_S2B (CD), Table_S2C (IBS), and Table_S2D (UC) for differential expression analysis of nutrition-related terms between years using Kruskal-Wallis

550 test. Only those terms are shown where the adjusted p-value (P_{adj}) < 0.05 . Mean expression
551 indicates the mean document-based normalised frequency obtained for a specific term for each
552 interval.

553

554

Figure 1(on next page)

Schematic of the keywords searched in PubMed search engine for the gastrointestinal diseases and nutrition-related literature

Twelve possibilities (4 X disease groups by 3 X nutritional categories) were searched in a 26-year timeline. The returned abstracts were grouped together in pairs of years and collated for a given group.

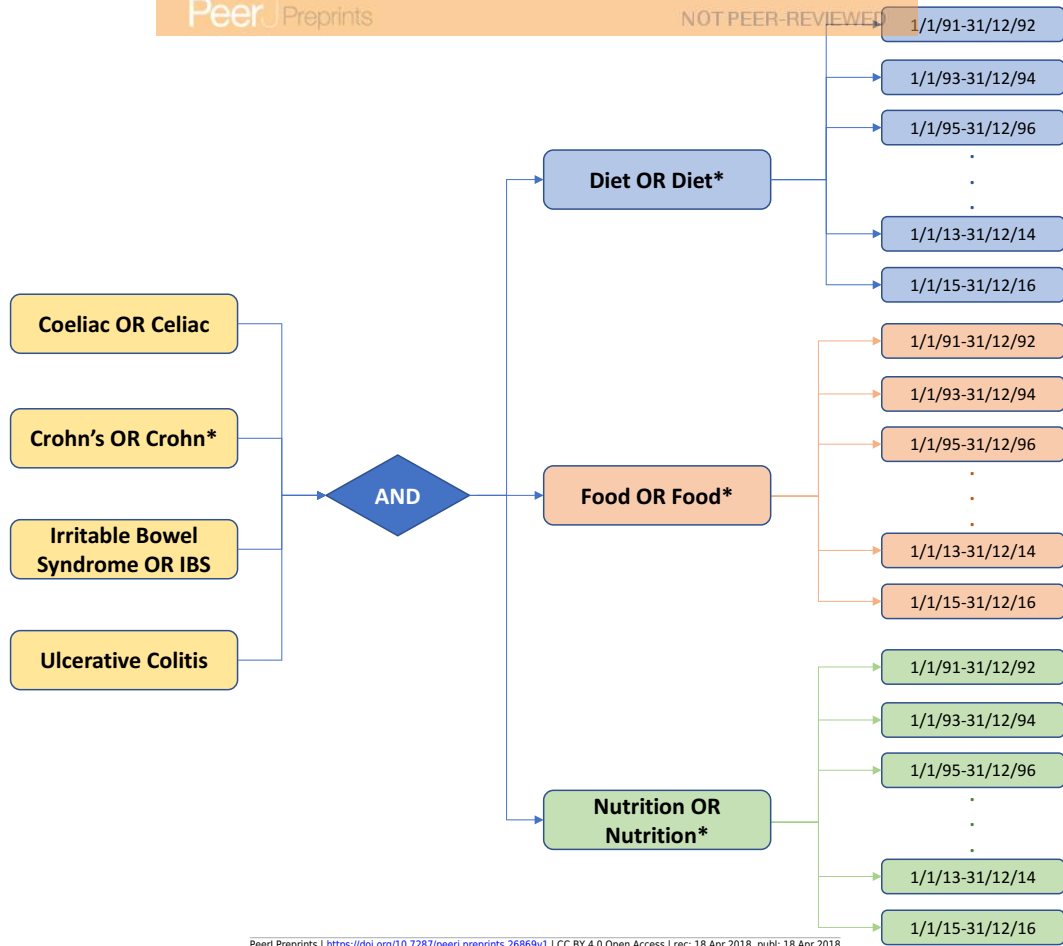
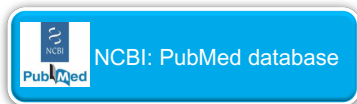


Figure 2(on next page)

Schematic of the workflow for the automated identification and analyses of ontological terms in literature data.

The abstracts returned from a keyword search in PubMed database are extracted and then processed with the pyTag workflow, where all the ontological terms are listed and annotated. After the annotation, a frequency table of the identified terms is generated and next subjected to statistical analysis.



Export Citations

Export BibTeX files

Extract PubMed IDs

Download abstracts

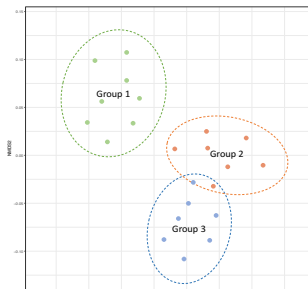
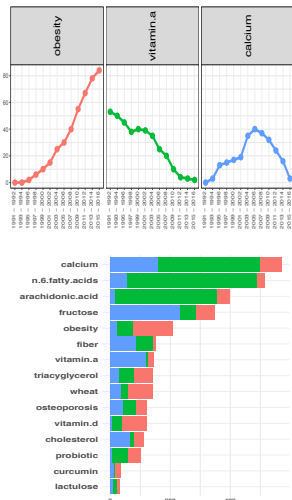
abstracts

EXTRACT Tagger

Generate table with the identified terms

Create frequency table

Multivariate Statistical Analysis



pyTag

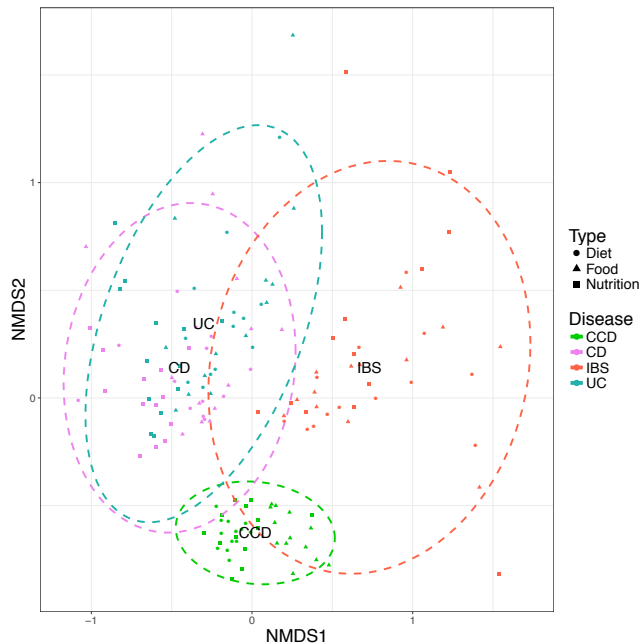
- NCBI Taxonomy
- Environment Ontology
- Disease Ontology and Mammalian Phenotype Ontology
- BRENDA Tissue Ontology
- GO Biological process
- GO Cellular component
- GO Molecular function
- Genes/Proteins (STING, RAIN)
- Small Molecule Compounds (STITCH)

Figure 3(on next page)

Non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance demonstrating clustering of IBD and non-IBD groups in the 26-year timeline.

Points indicate searches in pairs of years. A) Ellipses describe 95% CI of standard deviation for a given group. B) Dashed arrow represents transitions in the timeline. The size of the points corresponds to the date they describe where smaller size indicates earlier years and larger one more recent dates. SD=Starting date (1991-1992), ED=Ending Date (2015-2016), other=intermediate dates.

A



B

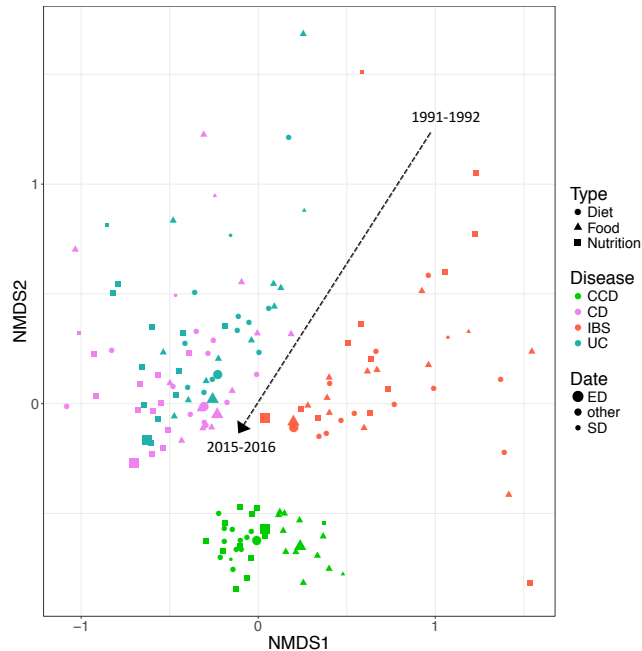


Figure 4(on next page)

The relative contributions to beta diversity (LCBD) per disease condition.

LCBD analysis demonstrating temporal variations in the literature of each disease group (distances from group average). Loess curve with shaded 95% CI illustrates the trends for each disease condition.

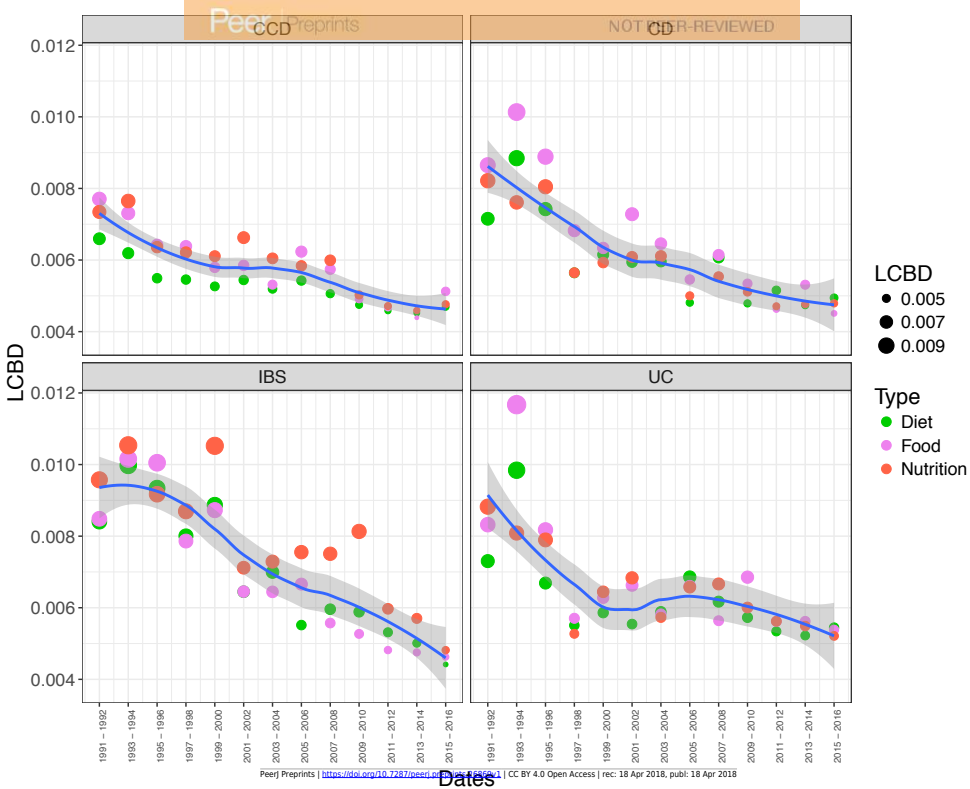


Figure 5(on next page)

Top 20 most frequent ontological terms in the literature of each disease condition for the entire time frame (1991-2016).

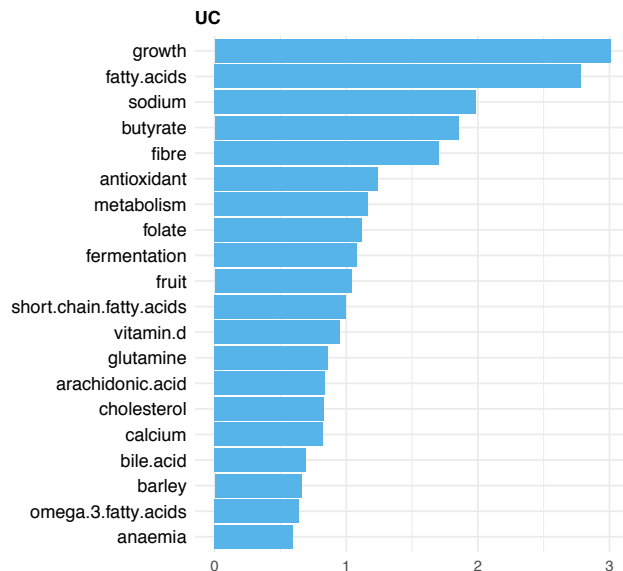
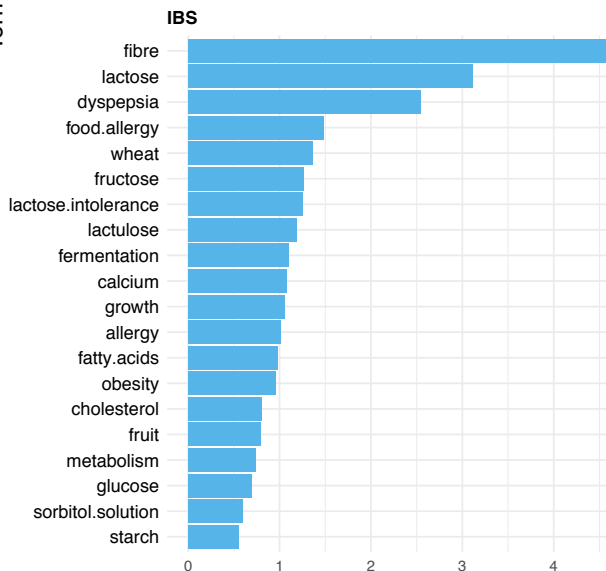
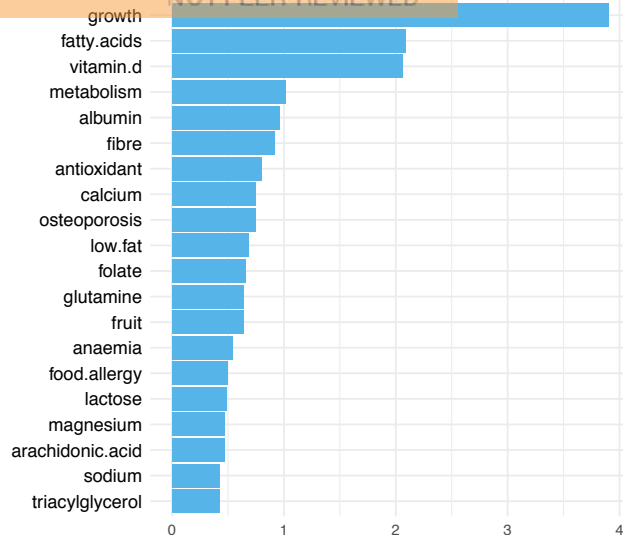
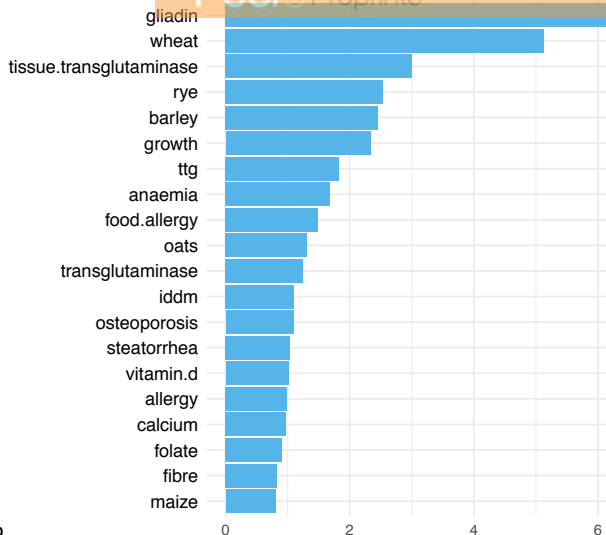
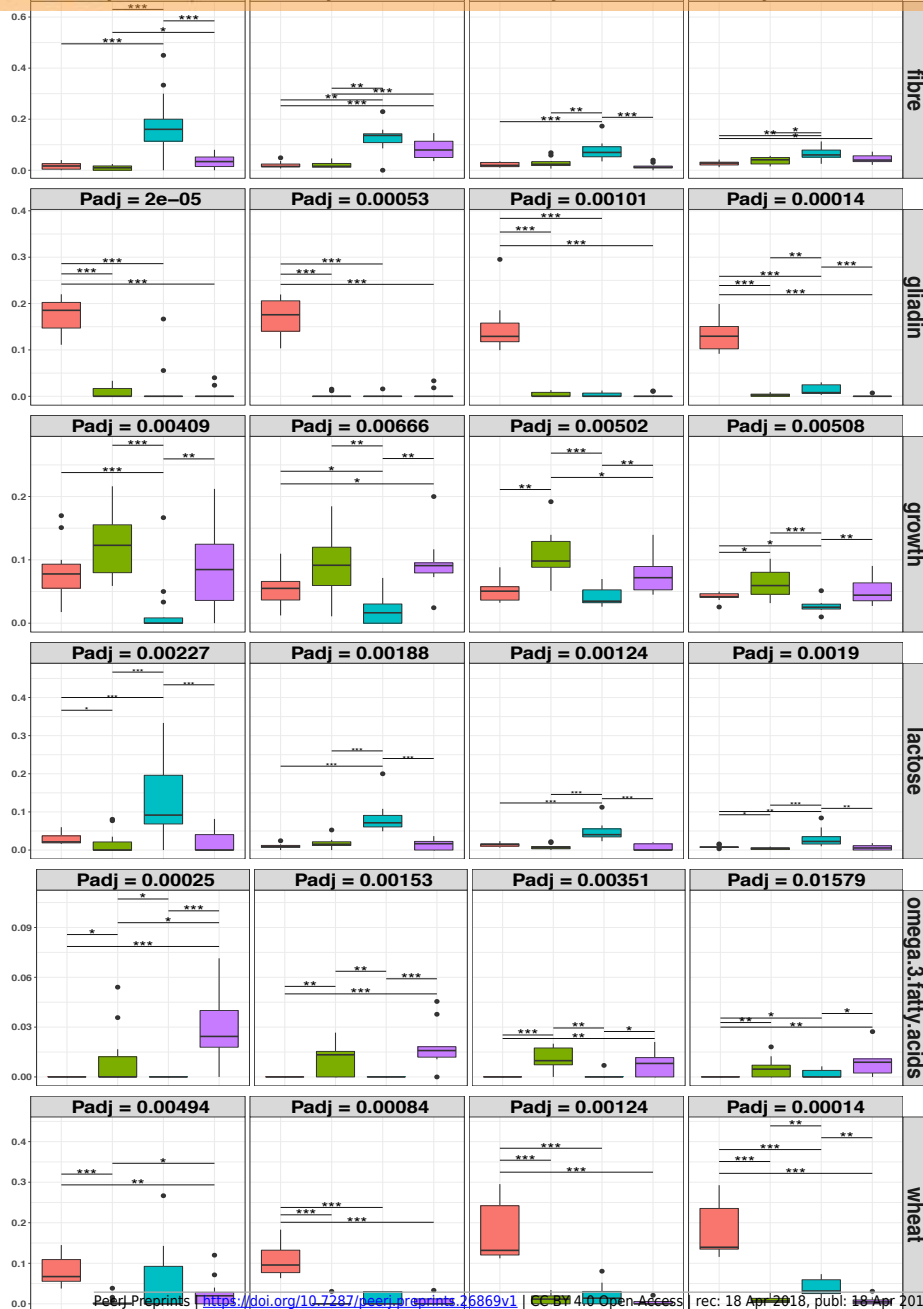


Figure 6(on next page)

Ontological terms whose frequency differentiated between the disease groups, over separate subsets of time intervals.

Box plots indicate the median, lower and upper quartiles of the document-based normalised frequency obtained for a specific term from the searches performed over the dates of a time interval, across the nutritional categories: Nutrition, Food and Diet, for a single group. Filled circles represent outliers. Dunn's comparison with asterisks indicating significant differences $*=p<0.05$, $**=p<0.01$ and $***=p<0.001$.



Group

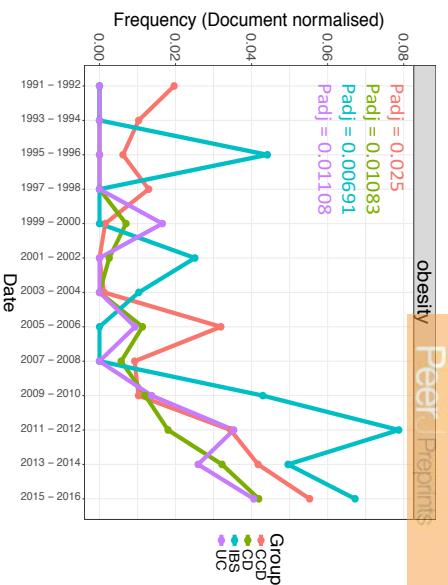
- CCD
- CD
- IBS
- UC

Figure 7 (on next page)

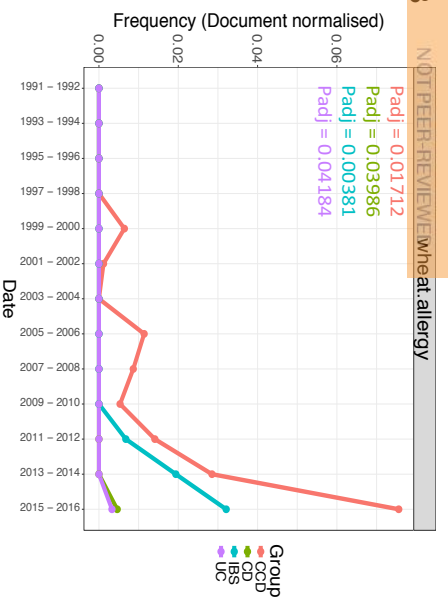
Trends of ontological terms whose frequency differentiated temporally in the literature of the gastrointestinal conditions.

Plots A) and B) describe the prevalence over time of *obesity* and *wheat allergy* respectively, in all disease groups, and plot C) describes the prevalence of terms found to differentiate over time in relation with CD and UC. Points indicate the mean document-based normalised frequency obtained for a specific term from a search conducted for a pair of years across the nutritional categories: Nutrition, Food and Diet, for a single disease group.

A



B



C

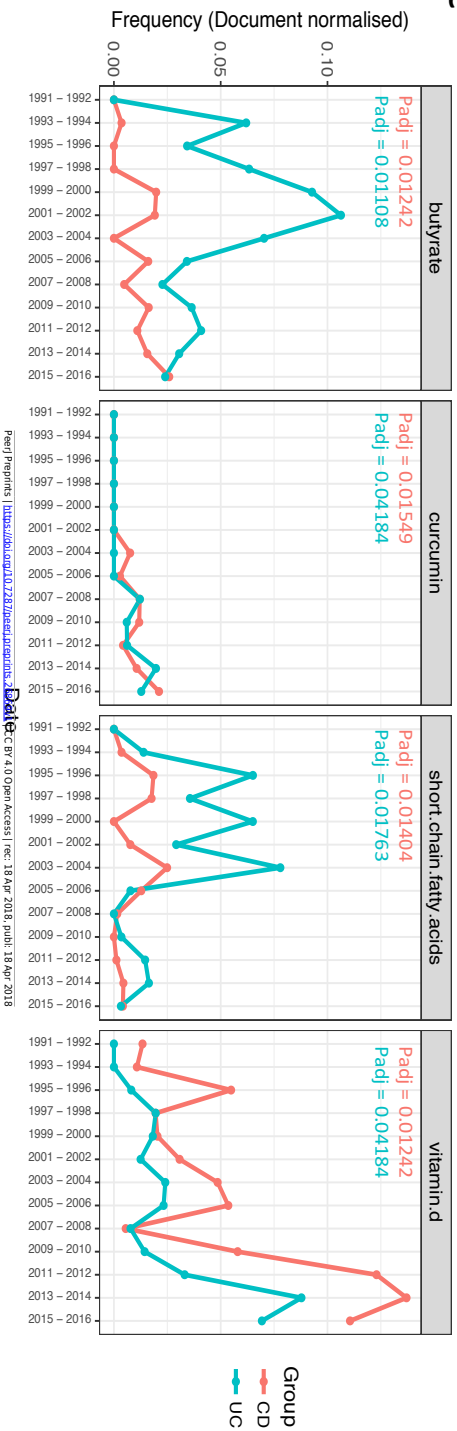


Table 1 (on next page)

Significance analyses performed on the identified ontological terms.

Temporally changing terms were explored for each disease group individually (Subset size). Ontological terms becoming significant between the groups were also explored using differential analysis in separate time intervals. An adjusted p-value (P_{adj}) < 0.05 was considered significant in each test. Percentage indicates the number of terms found significant over the size of the subset used for significance testing. n = total number of nutrition-related terms in the initial composite frequency table.

1

Number of ontological terms that differentiated over time in each disease condition			
Disease Group	n = 445	Significant terms (Padj < 0.05)	Percentage
	Subset size		
CCD	372	99	26.61 %
CD	385	185	48.05 %
IBS	287	169	58.88 %
UC	369	162	43.90 %
Number of ontological terms that differentiated between the disease conditions over separate time intervals			
Time interval	n = 445	Significant terms (Padj < 0.05)	Percentage
	Subset size		
1991 – 1998	290	51	17.58 %
1999 – 2004	337	73	21.66 %
2005 – 2010	383	62	16.18 %
2011 – 2016	425	143	33.64 %

2