Using patient-generated research questions to develop an ontology of Crohn's disease

Laura Christopherson, RENCI, University of North Carolina at Chapel Hill
David Borland, RENCI, University of North Carolina at Chapel Hill
Charles Schmitt, National Institute of Environmental Health Sciences

CCFA Partners

Crohn's disease is a chronic condition that affects every facet of patients' lives (e.g., social interaction, family, work, diet, sleep). Thus, treatment largely consists of disease management. The University of North Carolina chapter of the Crohn's and Colitis Foundation, CCFA Partners, has created an interactive website that—in addition to providing helpful information and disease management tools—offers a discussion forum for patients to discuss their experiences and suggest new lines of Crohn's research.

Our goal is to help CCFA researchers and physicians better understand how patients think about their condition and what research questions these patients would like the researchers to pursue.

The Data

97 posts (research questions with descriptions)

121 comments

17,322 words

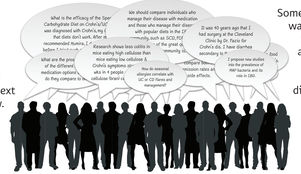
An Example Post

Question: *Nicotine has shown to be effective for UC in some individuals, both prior- and non-smokers. What is the mechanism? Does nicotine affect the microbiome, the immune system or both?*

Description: *Big Pharma will not take on the role of studying nicotine as there is no $$$ in it. Few studies with small sample sizes have been done but more research is needed.*

Why Visualization?

Sifting through all the questions and comments on the discussion forum is too time consuming. As the site gets used more and more, the text will continue to grow. Also, it is difficult to make sense out of so much text. Trying to do so is time and labor intensive.

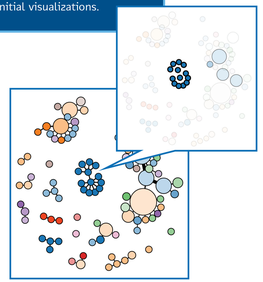


Something more efficient was needed to identify the common themes and determine which research questions are most frequently discussed by patients. Visualization offers a way to summarize the data and draw out the salient questions and needs of the patients.

Our Initial Approach

Words	Counts
ulcerative colitis	20
inflammatory bowel	13
controlled trial	10
inflammatory bowel disease	10
bowel disease	10
top priority	10
periodontal disease	9
disease activity	8
vitamin d	7
ibd management	5
other auto-immune	4
disease	
small bo	
combina	

Phrases were extracted from forum text (i.e., from questions, descriptions, and comments). Those phrases with >2 appearances were used to create the initial visualizations.



This didn't turn out to be as informative as we'd hoped. The frequency of words and/or word phrases didn't successfully capture the 'aboutness' of the conversation on the CCFA forum.

What does this mean, really?

I have heard from numerous people that when pregnant, they experience no symptoms of Crohns. I would like to know why this is and maybe we could find a medication that accompanies that.

Why an Ontology?

An ontology provides a richer representation of the information. Themes or topics become 'classes,' and more associations—beyond simply hierarchical and synonomous—are possible between classes (e.g., medication *manages* symptom). Ontologies are machine-readable; and the biomedical community actively and frequently uses ontologies (e.g., Gene Ontology, Human Phenotype Ontology) in everyday work.

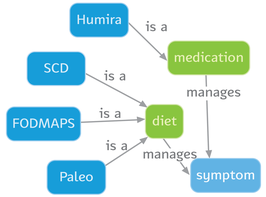
Look at examples A and B.
What are they really about?

A

We should compare individuals who manage their disease with medication and those who manage their disease with popular diets, such as SCD, FODMAPS, paleo, etc. Some people are able to manage their disease with with diet alone, but many take medication. Why do particular meds work for some, and particular diets work for others?

B

What is the efficacy of the Specific Carbohydrate Diet on Crohn's/UC? When I was diagnosed with Crohn's, my GI told me that diets don't work. After my GI recommended Humira, I decided to try SCD before I tried such a potent medication. The diet started working within days and, by the end of the first month, most of my symptoms were gone. I would like to know if SCD works for others and, if so, what percentage of patients.



Content Analysis

Manifest content is what you see in text, e.g., the occurrence of a particular word in a text. Our word frequencies are an example of identifying manifest content. Because we did not consider that to be effective in identifying the true 'aboutness' of CCFA forum conversation, we performed latent content analysis. "An example of latent content is the level of research anxiety present in user narratives about their experiences at the library" (Wildemuth, 2009). In other words, the user may not say exactly, "I am so anxious." Instead, their anxiety may be implied.

"Sometimes there is no existing theory or research on your message populations; you may not know what the important variables are. The only way to discover them is to explore the content (Wildemuth, 2009)." We read through all the questions, descriptions, and comments. For each post, we identified what appeared to be the main thrust of what the person was saying.

Recall the example:

I have heard from numerous people that when pregnant, they experience no symptoms of Crohns. I would like to know why this is and maybe we could find a medication that accompanies that.

On its face, this appears to be talking about pregnancy. But is it really? Could this person be talking about a physical state of being, specifically a hormonal state, that often coincides with a reduction in symptoms? The author provides an example of this state—pregnancy, but this state could manifest in other ways (e.g., birth control meds). In fact other conversants used birth contol as an example when discussing the same idea. What this person really wants to know is: can we make a medication that puts your body in a reduced-symptom state?

Themes that were assigned to this post included *hormonal state*, *medication*, *symptom*, and *remission*. These became classes in the ontology.

*Wildemuth, Barbara M. (2009). *Applications of Social Science Research Methods to Questions in Information and Library Science.* Libraries Unlimited.

The Ontology

337 classes

2261 annotations (e.g., concept definitions, labels)

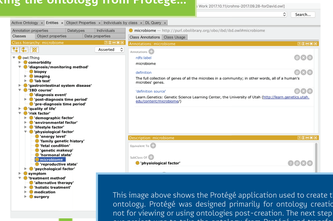
51 individuals

REPRESENTATIVE CLASSES	POSTS
comorbidity	11
diagnosis/monitoring method	7
IBD course	39
pre-diagnosis time period	1
diagnosis event	5
post-diagnosis time period	31
quality of life	8
risk factor	58
demographic factor	7
environmental factor	18
lifestyle factor	20
physiological factor	28
psychological factor	5
symptom	36
gastrointestinal manifestation	12
extra-gastrointestinal manifestation	3
treatment method	50
alternative therapy	7
holistic treatment	12
medication	28
surgery	13

The numbers in the POSTS column in the table above represent the number of questions/descriptions that discussed the class in the REPRESENTATIVE CLASSES column on the left. In some cases, you will note that the number of posts for a superclass (e.g., *symptom*) does not equal the sum of the posts for its subclasses (e.g., *gastrointestinal manifestation* and *extra-gastrointestinal manifestation*). This is because in many cases the patient discussed the superclass generally, i.e., discussed symptoms in general, rather than naming specific symptoms or types of symptoms.

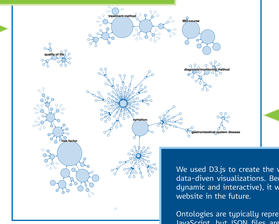
The Visualization

Taking the ontology from Protégé...



This image above shows the Protégé application used to create the ontology. Protégé was designed primarily for ontology creation, not for viewing or using ontologies post-creation. The next step in our project was to take the ontology from Protégé and transform it into a user-friendly visualization to help CCFA physicians better identify the key themes (i.e., the ontology classes) discussed in the CCFA forum, and to help demonstrate the prevalence of these themes as they appear in the research questions posed by CCFA patients.

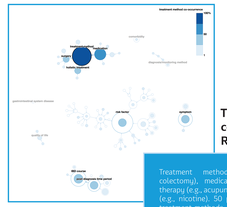
...to an interactive visualization!



Risk Factor

A risk factor is some sort of influence on a patient's condition. It can have a positive or negative effect. 58 posts touch on one or more risk factors. 28 of the 58 posts discuss a gene-related risk factor, e.g., the role of one's genetic makeup, one's microbiome, or concerns for passing the condition to a child.

This suggests that CCFA patients are keenly interested in the origins (genetics) of their disease but also in how their genes continue to impact their disease throughout their lives. This interest even extends to less commonly discussed gene-related topics such as the microbiome. (Has the everyday person on the street even heard of the term microbiome?) Is this concern about genes typical for patients of other chronic conditions? How often do patients of other chronic conditions delve into this level of detail (down to the microbiome) about their condition?



Treatment Method co-occurrence with Risk Factor

Treatment method includes surgery (e.g., colectomy), medication (e.g., Humira), holistic therapy (e.g., acupuncture), and alternative therapy (e.g., nicotine). 50 posts discussed one or more treatment methods.

19 of the 50 posts that discuss treatment method also discuss risk factor. For example, posts discuss wanting more research on how a particular risk factor affects the efficacy of a treatment method; or posts request the creation of a treatment method designed to respond to a particular risk factor.

6 of the 19 are what we call role-shifts, e.g., risk factors are discussed as treatment methods (diet, exercise), or a treatment method is discussed as a risk factor (the safety of biologic medications which have been linked to infections, and, worse, cancer). Currently a class's potential to shift roles is not modeled in the ontology, but it needs to be.

We used D3.js to create the visualization of the ontology. It is a JavaScript library designed for creating data-driven visualizations. Because it is based on JavaScript (a scripting language for making webpages dynamic and interactive), it will enable us to integrate the visualizations into the existing CCFA Partners website in the future.

Ontologies are typically represented in Web Ontology Language (OWL). OWL files are not easily read into JavaScript, but JSON files are. JavaScript Object Notation (or JSON) is a "lightweight data-interchange format" based on a subset of JavaScript. So to transform the OWL file into an interactive visualization using D3, we first needed to convert the OWL file into JSON. OBO Graphs—"a graph-oriented way of representing ontologies in … JSON format"—proved to be ideal for this.

Using D3, we implemented a custom ontology visualization that uses a force-directed network layout to show the ontological hierarchy. Each post was classified by one or more ontology terms. Those categorizations were loaded in CSV (comma-separated value) format, enabling the visualization of prominent terms, and the interactive highlighting of term co-occurrence.

D3: Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301-2309.
See also https://d3js.org
OWL: https://www.w3.org/TR/owl-features
JSON: https://www.json.org/
OBO Graphs: https://github.com/geneontology/obographs