Biology & Philosophy (2019) 34:61
https://doi.org/10.1007/s10539-019-9713-z

Why expect causation at all? A pessimistic parallel with neuroscience

Javier Gomez-Lavin[1]

Received: 17 July 2019 / Accepted: 10 October 2019 / Published online: 21 November 2019

Abstract
In their target article, Lynch, Parke, and O'Malley argue against the quick application of causal, interventionist explanatory frameworks to microbiomes and their purported role in many disparate states, from obesity to anxiety. I think the authors have undersold the force of their argument. A careful consideration of the scope of their claims, made easier by a parallel drawn from the history of explanation in neuroscience, yields a productive pessimism: that causal explanations likely operate at the wrong level of analysis for dynamic, distributed, *Quineian* entities like the microbiome. That is, we shouldn't expect causal explanations for microbiomes *at all*—and this includes the authors' own "microbiome success story" of *C. difficile*. Neuroscience, with its own computationally challenging, dynamic entity—the brain—may provide lessons for how to approach something like predictive control over the microbiome.

Keywords Microbiomes · *Clostridium difficile* · *Quineian* · Neuroimaging · Explanation in neuroscience · Multivariate analysis

Lynch, Parke and O'Malley marshal an elegant and simple set of cases that ought to caution us against the ready attribution of causal explanations to the microbiome. In the face of a prevailing trend towards sloppy and casual causal claims in microbiology, and many of its concomitant—omic sciences, the authors articulate a clear explanatory framework, which they hope might help suss out productive sites for future empirical intervention.

I'm largely convinced. However, I don't think the authors have embraced the real strength of the pessimism underwriting their argument, even for their own exemplar of *C. difficile,* which they term a "microbiome success story." Rather, the spirit of

This comment refers to the article available at https://doi.org/10.1007/s10539-019-9702-2.

✉ Javier Gomez-Lavin
jglavin@sas.upenn.edu

[1] University of Pennsylvania, Philadelphia, USA

their argument motivates a general skepticism to the application of causal explanations for microbiomes *at all*. Causal, interventionist frameworks operate at entirely the wrong level of analysis for systems-based, highly-dimensional entities that feature global properties such as the microbiome. Systems-based entities may be better understood by an increased sensitivity to what I'll term their *Quineian* features—borrowing here from Fodor (1983)—and the structuring relations between these global, multi-dimensional properties and other functional states of the system.

Here I'm driven by a stark parallel between the microbiome-narrative, with its excited pace of findings, and cognitive neuroscience and neuroimaging, with which I'm more familiar. Consider the similarities: The microbiome is composed of something like 40 trillion organisms organized into possibly 1000 "species-like groups" smeared out over a 10 m tube covering dozens of square meters across multiple anatomically distinct regions with different biochemical compositions. The brain has something like 100 billion neurons (alongside many more additional cells) each often featuring several hundred synaptic endings, categorized into many morphologically and functionally distinct types, located at different positions in the cerebral column, often grouped into distinctive regions (think 'Brodmann areas'). Both the brain and microbiome display differences across individuals, are implicated in disease states, and have homologues across other species. Both rely on low-frequency measurement tools that sample noisy, derivative products—with gen- and metabolomics on the one hand and a slow hemodynamic response measured by fMRI on the other, the list goes on. More relevant for us, both offer the difficult computational challenge of isolating their respective causally efficacious components. What I'm offering, then, are small lessons derived from the history of neuropsychology as a template to draw inspiration from when, or rather if, a thoroughgoing pessimism about microbiome-level causation takes hold.

Though there's always a risk of overgeneralizing when drawing up analogies, I'm surprised by the similarities among the respective explanatory genealogies of these two domains. The authors rightly critique the circular label of dysbiotic microbiomes (ms. 12), but fail to criticize the similarly deflationary notion of microbiome "enterotypes," or the three most "stable" compositions of microbial taxa found by Arumugam et al. (2011).[1] An emphasis on microbiome classification recalls parallel paradigms across the history of psychology: from Gall's phrenology to Sheldon's debunked "somatotype" theory (Rafter 2008). Key explanatory moves in the history of neuroscience from classification, to lesion studies, to a central focus on pathology are being replicated in microbiome research at double speed with the rise of "enterotyping," microbiota transplants on germ-free mice (ms. 13), and a focus on disease-states and their etiology. Such steps make intuitive sense. Testing Gall's lucky theoretical hunch that the brain must be functionally divisible, the early neurologist Flourens pioneered the technique of ablation, which in part motivated other physiologists to consider cortical lesions as possible explanations of various psychopathologies, leading to the discovery of Broca's, Wernicke's and other specialized

[1] Enterotypes might turn out to be a twenty-first century star-chart, with many internet services helping you "find your enterotype" and matching you to a probiotic blend that they conveniently stock.

brain-regions (Pearce 2009). Using disease states as windows into the workings of these systems is certainly useful, but such an emphasis on pathology may lead us to overlook the constant complex dynamics that underlie their outsized everyday role. Just as the brain has no single "resting state," the microbiome must also have its hands constantly full (Klein 2014).

These moves are aimed at reducing the problem space posed by these vast, distributed, and dynamic entities as researchers attempt to identify and isolate their causal components. However, the computational problem posed by these systems is near intractable. Returning to the brain, consider that for all combinations of a binary eight-by-eight arrangement of pixels the amount of time needed to determine how a given neural population encodes the possible images, assuming you could sample the region at one image per second, would take orders of magnitude longer than the history of the universe. Given the number of possible neural regions and the number of possible stimuli and it's easy to see how a brute force mapping between the two lies far beyond our reach (for more on this problem, consult Haynes 2015).

The problem is exacerbated by how neural regions actually encode information. Barlow's (1953) experiments on the frog retina and Hubel and Wiesel's vertical-slit experiments (1962) on the cat retina each showed how specific neurons could be *tuned* for perceptible properties, such as position or orientation. The idea that neurons encode specific stimulus properties by varying their firing rate was cemented as a kind of *univariate dogma* that underwrote the excitement surrounding neurophysiology in the twentieth century (Postle 2015). All that was needed was to decode the tuning properties of neurons and we'd have a ready read-out of how the brain processes information, or, as Barlow put it: "a description of that activity of a single nerve cell… is a complete enough description for a functional understanding of the nervous system" (1972, p. 380).

Of course, things turned out to be far more complicated, as neurons are capable of flexibly retuning their firing rates, not to mention that firing rates are a symptom of many deeper cellular mechanisms (Duncan 2001). In neuroimaging, a similar turn towards *multivariate* models has occurred in the last decade (Haynes 2009, 2015). Rather than attempt to identify which brain regions are the cause of some psychological process by observing their relative blood-oxygenation levels during a task as an analog for their level of "activity," multivariate approaches aim to unveil how mental contents are encoded as distributed patterns of activity across the cortex. Caricaturing things a bit, imagine that you're trying to determine how the brain processes images of cats. Fifteen years ago you would place a subject in a scanner, show them alternating pictures of cats and a contrast class, say houses, and then you'd subtract the smoothed and regionalized hemodynamic responses across the two classes. If you were lucky, you'd find one region with an activation profile that responded preferentially to cat-images. This region would be baptized the "cat-area" and you'd move on to other stimuli. Inferentially, one can't draw too many conclusions from such an approach, as it's not even clear what activity in this area is *doing*—is it encoding features of cats, or performing some more epiphenomenal role? With multivariate techniques, you'd start similarly, by gathering hemodynamic data on images of cats and houses, and then you'd use this data to train a classifier—often a simple linear decision boundary. The classifier can then be used to predict

whether a subject, in a later test, saw images of cats or houses, in part by comparing the cross-cortical patterns of activity with those from prior labelled runs.

This move hints at a new sensitivity towards the *Quineian,* distributed properties of the brain. Just as there is no true "resting state" of the brain, it's clear that no single region of the brain works in silence; just because you've found a region of the brain tuned for images of cats, it only does its job *because* it is coupled with the rest of the system. And as multivariate techniques can attest, we can often find stimulus-related activity *throughout* the entire brain. Now, this doesn't solve the computational problem introduced above—in fact, it sharpens it. So, is there a solution to be found? Perhaps, but it will largely depend on the clever application of these machine learning techniques, including the creation of models that try to predict which stimuli trigger a given neural response and vice versa.[2] Whether you consider this a solution will likely depend on how willing you are to give up a thorough, mechanistic mapping of brain to behavior—the kind of mapping that may yield sites for precise causal intervention and control. Insofar as the brain genuinely gives us a good template to follow here, then we must acknowledge the pragmatics entailed by the move from causal to predictive, associative paradigms in the face of these distributed, dynamic, and dense systems and the computational challenges they foist.

But even if you don't buy my larger parallel, the authors could go further still within the constraints of their own examples, as their *C. difficile* case fails to fit the causal, interventionist framework they've set out.

The authors tout the highly publicized use of faecal microbial transplant (hereafter 'FMT') to cure patients suffering from intestinal colitis associated with *C. difficile* as "a microbiome success story that not only produces a treatment but also leads back to Koch's postulates" (ms. 19). I want to push back on their interpretation of this narrative, and on its coherent connection to the postulates. Instead we should appreciate how this case may better fit a *Quineian* interpretation of system-level entities and their functional properties.

The authors highlight the successful treatment of *C. difficile* by FMTs as an "example of microbiome research [that] provides a good casual explanation," as recent research has revealed a number of key components that may play outsized roles in curing *C. difficile* infection. *C. difficile* allows the authors to counter a *prima facie* microbiome-level explanation, wherein transplanting an entire microbiome by FMT results in a 90% or greater "cure rate" of patients suffering from *C. difficile* induced colitis (Bakken et al. 2011). Relying on research by Stein et al. (2013) and Buffie et al. (2015), the authors conclude that a more proportional and specific explanation—that is, an explanation better suited to an interventionist framework—is within reach as this research isolates "a restricted number of bacterial components [that] explain *C. difficile* cures." That is, contra the authors' initial characterization of *C. difficile* treatment as *microbiome* success story, they ultimately attribute

[2] For instance, you could train a decoding model to predict, given a pattern of neural activation, which stimulus was most likely present. Likewise, you could train an encoding model to do the opposite. Applying statistical learning techniques, including Bayesian strategies, can further the predictive power of these models, as Schoenmakers et al. (2013) demonstrate.

any success to a deflationary scheme wherein only a few key players should get the credit.

There are at least two reasons to push back on the authors' deflationary move. First, the research cited does not describe a mechanism nor does it provide a causal sketch for how FMTs *in particular* cure *C. difficile*. Rather, it describes how changes to microbiome populations and metabolic pathways may *allow* for pathogenic *C. difficile* colonization. In particular, Stein et al. (2013) describe a "mechanism" where antibiotic administration inhibits populations of native *Blautia* and *Akkermansia* bacteria, which in turn allow *Enterococcus* populations to grow "facilitating colonization by *C. difficile*" (p. 6, consult their Figure 5). Likewise, Buffie et al. (2015) examine how increased *C. scindens* populations confer resistance to *C. difficile* infections via secondary bile acid synthesis, a major metabolic function of these bacteria (p. 207). Both sets of authors do allow that their mechanisms of choice may have a role to play in FMTs, but such a role is not studied.

This leads to a second reason to resist the target article's example: the multiple realizability of microbiome-dependent functions. *Resistance* to *C. difficile* colonization and its ultimate pathogenesis is likely conferred by many overlapping mechanisms, as evidenced in the cited studies. Likewise how FMTs *cure C. difficile* colonization is certain to be similarly mediated. These are microbiome-level *functional* descriptions.

Furthermore, it is not clear how this example connects to Koch's postulates: after all, the authors are not focused on whether *C. difficile* causes colitis, but whether FMTs "cause" a cure of *C. difficile*. However this seems to be a misapplication of the postulates.

So how should we interpret FMT treatments of *C. difficile*? I would argue, instead, that we should treat this as a systematic change in the global properties of a microbiome: moving from one microbiome with locally high concentrations of *C. difficile* in the large intestine, to one without. As shown by the cited research there are likely *many* paths to such a change, likely overwhelming any one interventionist scheme. The functional ramifications of such a change, for the host at least, involve a reduction of TcdA and TcdB intoxication of epithelial cells associated with lesioning and the ultimate person-level disease state (Voth and Ballard 2005, Fiorentini et al. 1998). Clearly, causal explanations subject to intervention and manipulation—as in the case of *C. difficile* mediated TcdA and TcdB production—have a role to play; but their role will not exhaust the dynamic, global, and functional properties of such large scale, *Quineian*, systems.

References

Arumugam M, Raes J et al (2011) Enterotypes of the human gut microbiome. Nature 473:174–180

Bakken JS, Borody T et al (2011) Treating *Clostridium difficile* infection with fecal microbiota transplantation. Clin Gastroenterol Hepatol 9:1044–1049

Barlow HB (1953) Summation and inhibition in the frog's retina. J Physiol 119:69–88

Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology. Perception 1:371–394

Buffie CG, Bucci V et al (2015) Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. Nature 512:205–208
Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. Nat Rev Neurosci 2:820–829
Fiorentini C, Fabbri A et al (1998) *Clostridium difficile* toxin B induces apoptosis in intestinal cultured cells. Infect Immun 66(6):2660–2665
Fodor J (1983) The modularity of mind. MIT Press, Cambridge
Haynes J (2009) Decoding visual consciousness from human brain signals. Trends Cognit Sci 13(5):194–202
Haynes J (2015) A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron 87:257–270
Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 166:106–154
Klein C (2014) The brain at rest: what is it doing and why it matters. Philos Sci 81:974–985
Pearce JMS (2009) Marie–Jean–Peire Flourens (1794–1867) and cortical localization. Eur Neurol 61(5):311–314
Postle BR (2015) Neural bases of the short-term retention of visual information. In: Jolicoeur P, Lefebvre C, Martinez-Trujillo J (eds) Mechanisms of sensory working memory: attention and performance XXV. Academic Press, London, pp 43–58
Rafter N (2008) Somatotyping, antimodernism, and the production of criminological knowledge. Criminology 45(4):805–833
Schoenmakers S, Barth M, Heskes T, Van Gerven M (2013) Linear reconstruction of perceived images from human brain activity. Neuroimage 83:951–961
Stein RR, Bucci V et al (2013) Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol 9(12):1–11
Voth DE, Ballard JD (2005) *Clostridium difficile* toxins: mec of action and role in disease. Clin Microbiol Rev 18(2):247–263