Diversity in citations to a single study

Rhodri Ivor Leng (2021)

June 16, 2021

Background

- Diet heart hypothesis gained traction in the 50s
- Paul et al. 1963 reported findings that did not support the hypothesis
    - + many other findings
- This paper is highly cited (446 citations before 1985), uncommon for a negative findings paper
- **Question:** How did others interact with this paper?

Summary (Paul)

- The *Western Electric Study* by Paul et al. (1963) was examined
- Sample: 1,989 men, 40-55 years old; follow-up after nearly 4.5 years
- Which factors influenced who got CHD? (n=88)

“No Association”	**“Significant Relationship”**
Dietary variables (fats, carbs, protein, salt)	Elevated Serum Cholesterol
Alcohol Consumption	High Blood Pressure
Job Type (Psychosocial)	Coffee Intake
Job Related Physical Activity	Smoking Cigarettes

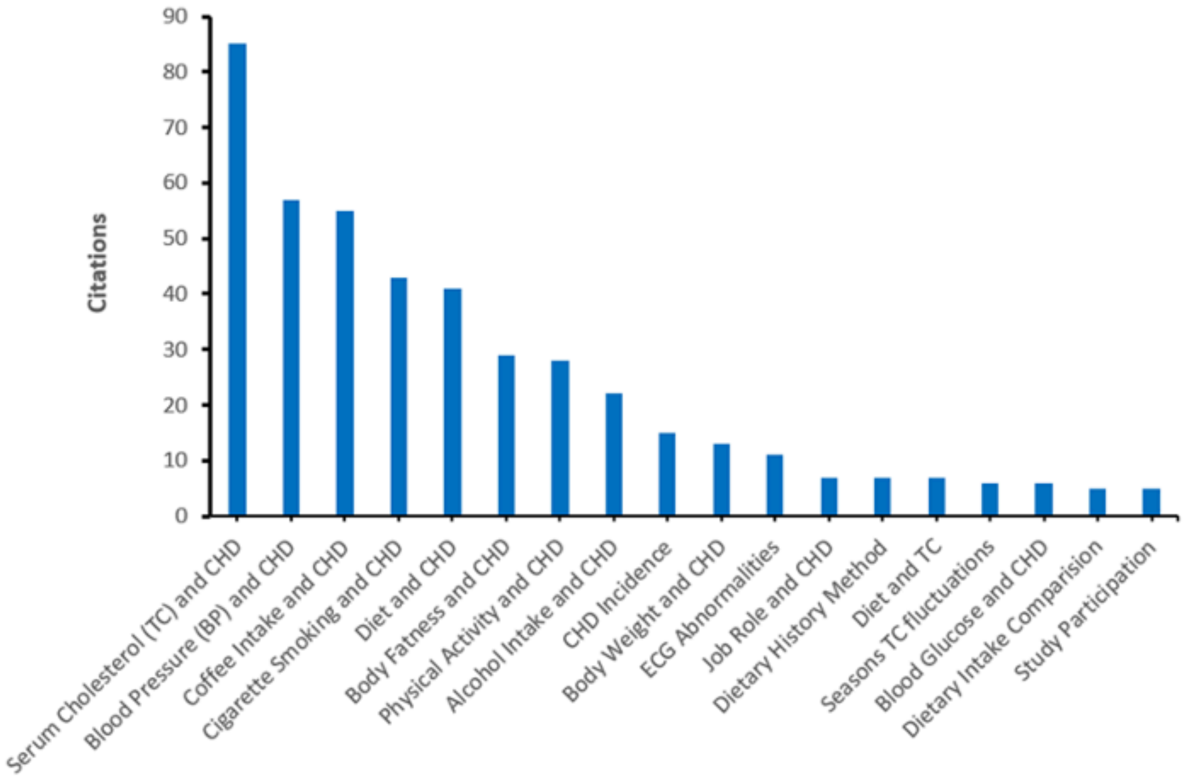
“Body Fatness” was also a factor mentioned in the paper, but it’s not mentioned which category it fell into.

Methods

- Paul et al. (1963) was cited 446 times between its publication and 1985
- 343 papers were accessible
    - In-text citations copied and analyzed
    - Papers could cite multiple findings
- Titles were used to create categories:
    - Explanans = “cause”
    - Explanandum = “effect”

Results

What findings were cited or cited together?



Results

What findings were cited or cited together?

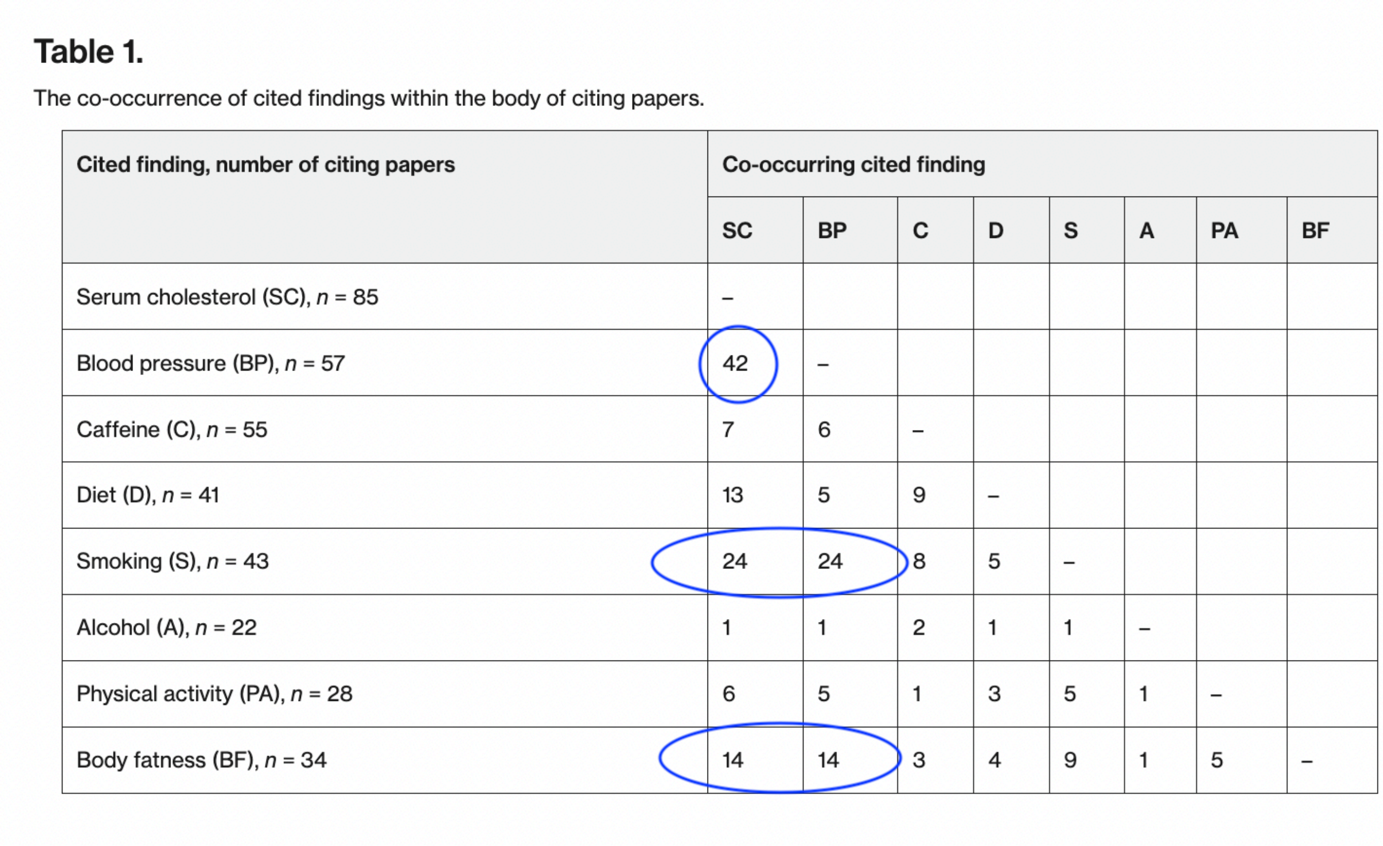


Table 1.

The co-occurrence of cited findings within the body of citing papers.

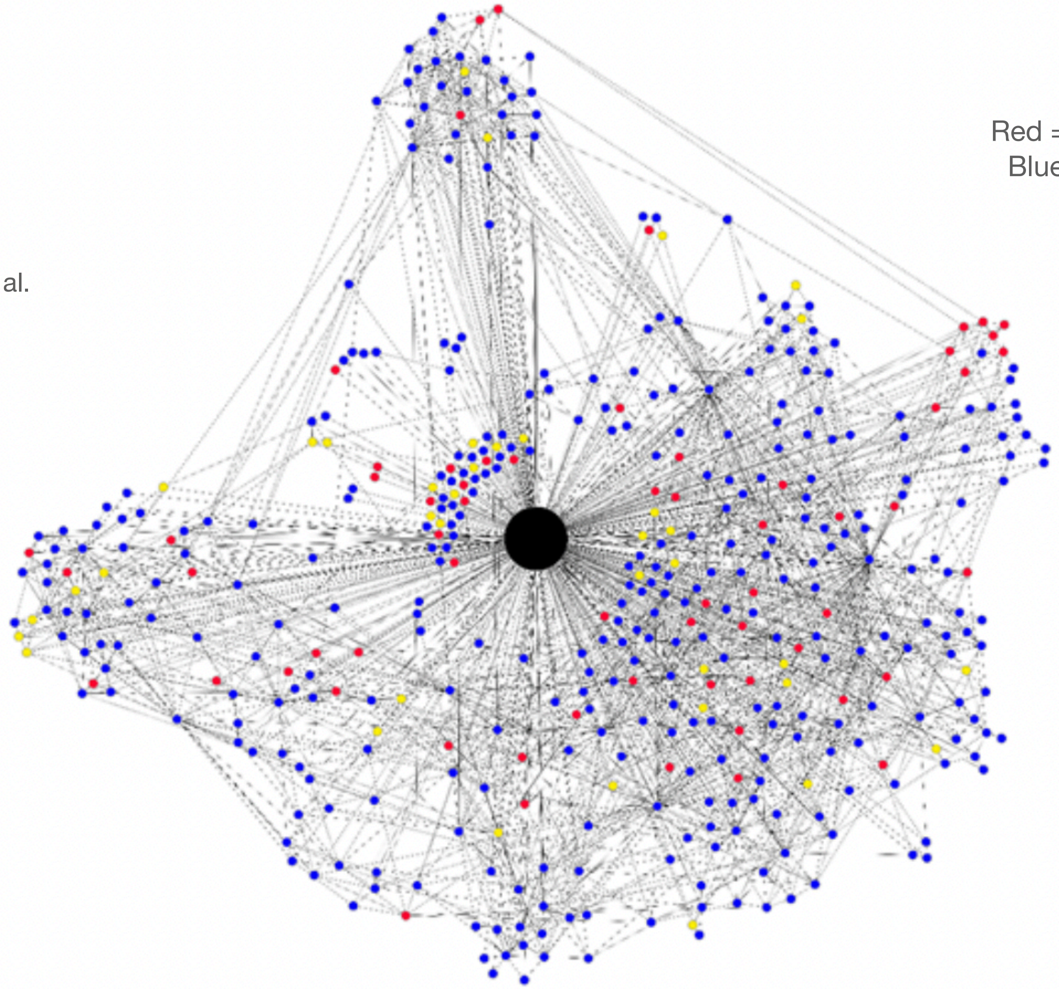
Cited finding, number of citing papers	Co-occurring cited finding							
	SC	BP	C	D	S	A	PA	BF
Serum cholesterol (SC), *n* = 85	–							
Blood pressure (BP), *n* = 57	42	–						
Caffeine (C), *n* = 55	7	6	–					
Diet (D), *n* = 41	13	5	9	–				
Smoking (S), *n* = 43	24	24	8	5	–			
Alcohol (A), *n* = 22	1	1	2	1	1	–		
Physical activity (PA), *n* = 28	6	5	1	3	5	1	–	
Body fatness (BF), *n* = 34	14	14	3	4	9	1	5	–

75% of papers mentioned one finding
25% mentioned more than one
If more than one finding mentioned, at
ast one of them is likely serum
cholesterol or blood pressure

Results

- Explanans and Explanandum did not provide enough information independently, many papers were missing one or the other
    - Leng conflated the two levels
- Papers fell in one of 10 categories:
    - General CHD
    - Serum Cholesterol
    - Diet
    - Caffeine
    - Blood Pressure
    - Psychosocial
    - Alcohol
    - Body Fatness
    - Physical Activity
    - Smoking

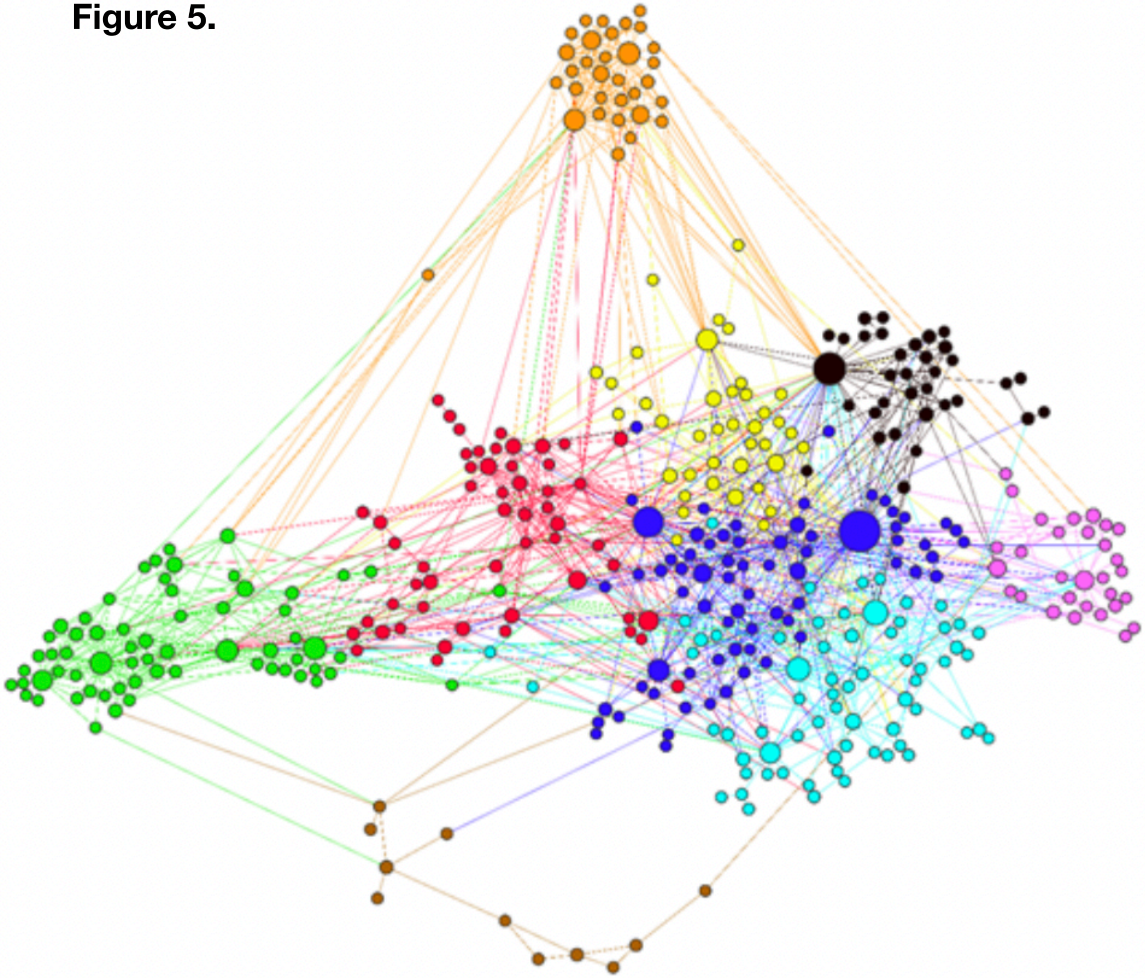
Figure 4.



Yellow = Not in English
Red = Not available (not scanned)
Blue = In English, text available

Node = paper that cites Paul et al.
Edge = direct citation

Figure 5.



Cluster	Color	#Nodes/#full data papers	% of cluster papers with common title category
Cluster 0	Blue	70/52	40% Unspecified CHD 37% Serum cholesterol
Cluster 1	Light blue	66/49	45% Unspecified CHD 22% Blood pressure
Cluster 2	Green	64/48	69% Caffeine
Cluster 3	Red	53/47	77% Diet
Cluster 4	Yellow	39/36	39% Unspecified CHD 28% Body fatness
Cluster 5	Pink	34/28	68% Physical activity
Cluster 6	Orange	32/27	93% Alcohol
Cluster 7	Black	30/21	71% Psychosocial
Cluster 8	Brown	11/8	88% Smoking

Larger nodes have been cited more frequently by other papers in the network
The color of the edge is the color of the citing paper

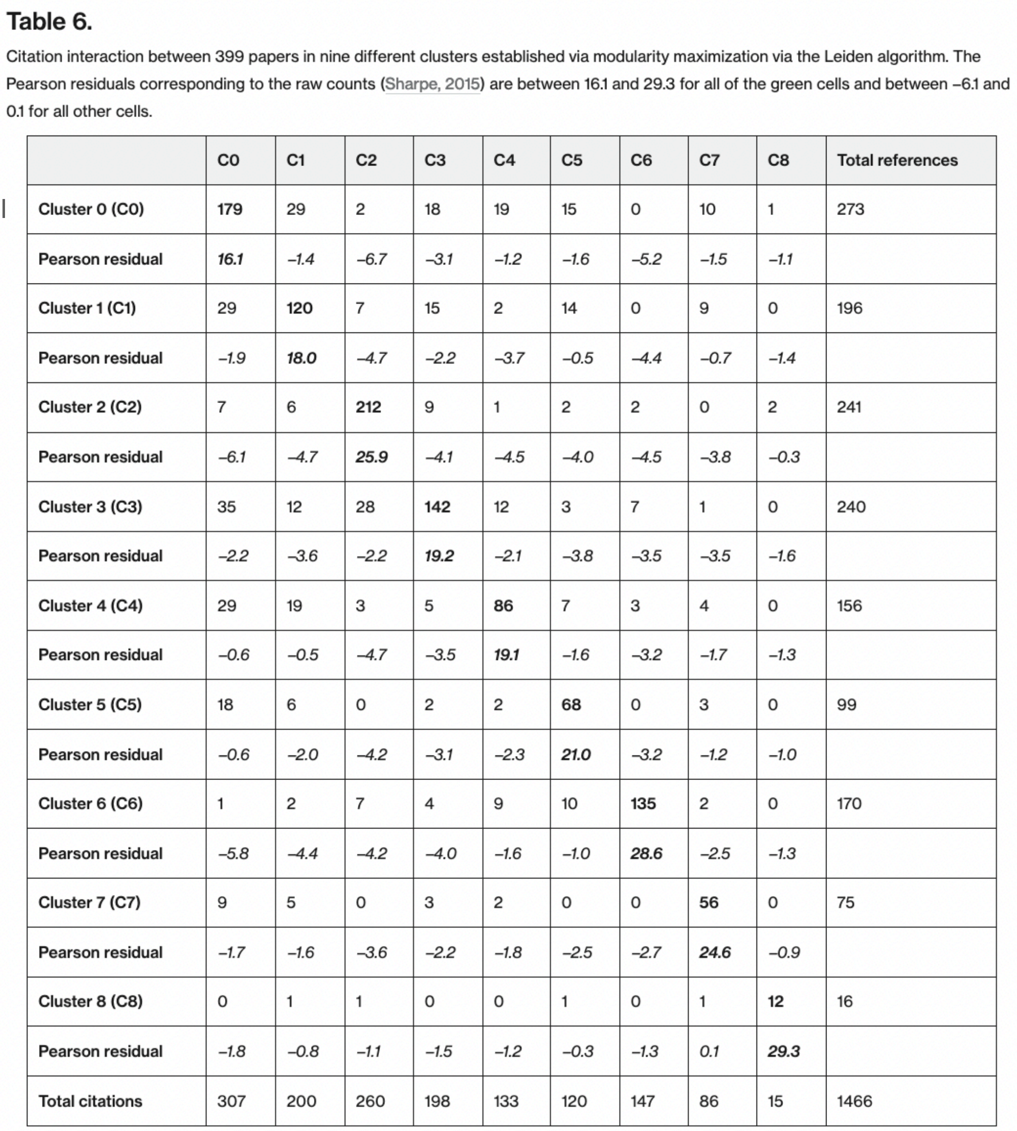


Table 6.

Citation interaction between 399 papers in nine different clusters established via modularity maximization via the Leiden algorithm. The Pearson residuals corresponding to the raw counts (Sharpe, 2015) are between 16.1 and 29.3 for all of the green cells and between –6.1 and 0.1 for all other cells.

		C0	C1	C2	C3	C4	C5	C6	C7	C8	Total references	% Inward-facing citations
Serum Cholesterol	**Cluster 0 (C0)**	**179**	29	2	18	19	15	0	10	1	273	66%
	Pearson residual	***16.1***	*–1.4*	*–6.7*	*–3.1*	*–1.2*	*–1.6*	*–5.2*	*–1.5*	*–1.1*		
Blood Pressure	**Cluster 1 (C1)**	29	**120**	7	15	2	14	0	9	0	196	61%
	Pearson residual	*–1.9*	***18.0***	*–4.7*	*–2.2*	*–3.7*	*–0.5*	*–4.4*	*–0.7*	*–1.4*		
Caffeine	**Cluster 2 (C2)**	7	6	**212**	9	1	2	2	0	2	241	88%
	Pearson residual	*–6.1*	*–4.7*	***25.9***	*–4.1*	*–4.5*	*–4.0*	*–4.5*	*–3.8*	*–0.3*		
Diet	**Cluster 3 (C3)**	35	12	28	**142**	12	3	7	1	0	240	59%
	Pearson residual	*–2.2*	*–3.6*	*–2.2*	***19.2***	*–2.1*	*–3.8*	*–3.5*	*–3.5*	*–1.6*		
Body Fat	**Cluster 4 (C4)**	29	19	3	5	**86**	7	3	4	0	156	55%
	Pearson residual	*–0.6*	*–0.5*	*–4.7*	*–3.5*	***19.1***	*–1.6*	*–3.2*	*–1.7*	*–1.3*		
Physical Activity	**Cluster 5 (C5)**	18	6	0	2	2	**68**	0	3	0	99	68%
	Pearson residual	*–0.6*	*–2.0*	*–4.2*	*–3.1*	*–2.3*	***21.0***	*–3.2*	*–1.2*	*–1.0*		
Alcohol	**Cluster 6 (C6)**	1	2	7	4	9	10	**135**	2	0	170	68%
	Pearson residual	*–5.8*	*–4.4*	*–4.2*	*–4.0*	*–1.6*	*–1.0*	***28.6***	*–2.5*	*–1.3*		
Psychosocial	**Cluster 7 (C7)**	9	5	0	3	2	0	0	**56**	0	75	75%
	Pearson residual	*–1.7*	*–1.6*	*–3.6*	*–2.2*	*–1.8*	*–2.5*	*–2.7*	***24.6***	*–0.9*		
Smoking	**Cluster 8 (C8)**	0	1	1	0	0	1	0	1	**12**	16	75%
	Pearson residual	*–1.8*	*–0.8*	*–1.1*	*–1.5*	*–1.2*	*–0.3*	*–1.3*	*0.1*	***29.3***		
	Total citations	307	200	260	198	133	120	147	86	15	1466	

How were findings used?

- Serum Cholesterol, Blood Pressure
  - This was an established finding. Others used Paul et al. as further evidence of the connection
- Smoking
  - Not an established finding, but Paul et al. was treated as a paper establishing the connection as fact
- Diet and Heart Disease
  - 41 papers reported negative findings
  - Others critiqued the methods - *Is* there a low-fat American diet?
- Coffee and Heart Disease
  - Exploration of the connection - Sugar? Confounding variables?

Conclusions

- Cherry-picking of convenient findings
- Diet-heart hypothesis community largely ignored these findings
  - The sugar-heart hypothesis community took it up as evidence of their hypothesis
    - Paul et al. did not mention sugar in their findings
- Impact had breadth but not depth

Critiques

- Conflation of explanans and explanandum
- Combination of general and specific factors
    - “Diet” is more general than “caffeine”
- Clustering methods = maximizing cluster size
    - Might not be best fit; how many singletons?
- Minor point = yellow a poor color choice in examining edges