Data in Brief 35 (2021) 106841

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Studying attitudes towards vaccine hesitance and California law SB 277 in online discourse: A dataset and methodology

Kali DeDominicis[a,*], Alison M. Buttenheim[b], Amanda C. Howa[c], Paul L. Delamater[d], Daniel Salmon[e], Nicola P. Klein[f], Saad B. Omer[g]

[a] *Department of Anthropology, Sociology, and Criminal Justice, Arcadia University, 450 Easton Rd., Glenside, PA 19038, United States*
[b] *Department of Family and Community Health, University of Pennsylvania School of Nursing, Center for Health Incentives and Behavioral Economics, Perelman School of Medicine, University of Pennsylvania, Claire M. Fagin Hall, 418 Curie Boulevard, Philadelphia, PA 19104, United States*
[c] *Hubert Department of Global Health, Rollins School of Public Health, Emory University, 201 Dowman Drive, Atlanta 30322, Georgia*
[d] *Department of Geography and Carolina Population Center, University of North Carolina at Chapel Hill, 123 Franklin St., Room 2152, Chapel Hill, NC 27516, United States*
[e] *Department of International Health and Health Behavior and Society, Bloomberg School of Public Health, Johns Hopkins University, Institute for Vaccine Safety, Bloomberg School of Public Health, Johns Hopkins University, 615N. Wolfe St., Room W5035, Baltimore, MA 21205, United States*
[f] *Kaiser Permanente Vaccine Study Center, 2000 Broadway, Oakland, CA 94612, United States*
[g] *Yale Institute for Global Health, Associate Dean (Global Health Research), Yale School of Medicine, Professor of Medicine (Infectious Diseases), Yale School of Medicine, Susan Dwight Bliss Professor of Epidemiology of Microbial Diseases, Yale School of Public Health, 1 Church St., New Haven, CT 06510, United States*

ARTICLE INFO

Article history:
Received 9 September 2020
Revised 1 February 2021
Accepted 2 February 2021
Available online 24 February 2021

ABSTRACT

This article presents data that are further analyzed and interpreted in "Shouting at Each Other into the Void: A Semantic Network Analysis of Vaccine Hesitance and Support in Online Discourse Regarding California Law SB277" [1].
This research modified snowball sampling, a technique usually used to generate chains of informants that illuminate the structure of social networks, to collect digital documents following a chain of web links and recommendations, thus

DOI of original article: 10.1016/j.socscimed.2020.113216
* Corresponding author.
E-mail addresses: kali.dedominicis@gmail.com (K. DeDominicis), abut@nursing.upenn.edu (A.M. Buttenheim), Amandahowa16@gmail.com (A.C. Howa), pld@email.unc.edu (P.L. Delamater), dsalmon1@jhu.edu (D. Salmon), nicola.klein@kp.org (N.P. Klein), saad.omer@yale.edu (S.B. Omer).

https://doi.org/10.1016/j.dib.2021.106841

Keywords:
Anti-vaccination movement
California SB 277
Digital methodology
Digital snowball sampling
Online discourse
Semantic network analysis
Vaccine hesitance
Vaccine policy

illuminating the underlying social, technical, and linguistic structure of online discourse. The resulting documents were manually coded according to the attitude towards vaccines they represented and/or the position they took with regard to California Senate Bill 277, a vaccine mandate policy that banned all nonmedical exemptions from school immunization requirements. Each attitude category, as well as the dataset as a whole, was subjected to quantitative linguistic analysis to identify key words and phrases in the data according to the frequency with which they appeared. A combination of that technique and semantic network analysis were used to generate clusters of related words that could be used for qualitative and narrative analysis, as detailed in the companion paper. The data collection and analysis processes described here will be of use to researchers conducting mixed-method analysis of online discourse who want their data to reflect the potential information and digital resources available to individuals who attempt to inform themselves about a particular topic using Internet searches. The data presented here could be useful for anyone seeking deeper insight into the linguistic and narrative patterns surrounding online debates about vaccination, controversial government policies, or both.

Specifications Table

Subject	Social Sciences (General)
Specific subject area	Attitudes towards vaccination, attitudes towards vaccine policy mandates, semantic network analysis
Type of data	Tables
How data were acquired	Modified snowball sampling, to reflect networks of web links rather than informants
Data format	Raw Analyzed
Parameters for data collection	Online documents were collected that discussed California Senate Bill 277 between June 2014 and June 2017
Description of data collection	Data were collected using a modified form of snowball sampling that illuminated the technological and social networks of public web links that individuals can access and participate in using a basic web search. It is comprised of an Excel spreadsheet that compiles the word frequency charts created and exported using NVivo during our study of attitudes towards Califonia law SB277, a vaccine mandate. There are eight total sheets representing the total dataset and then that dataset divided into three attitude categories ("vaccine supportive," "vaccine hesitant," and "neutral"). Each of those categories had two outputs, "stemmed" and "unstemmed." The latter compiled the raw lists of words in each category by frequency of use, while the former grouped words that began with the same collection of letters (e.g., "child" and "children").
Data source location	United States
Data accessibility	Data are included in this article and are uploaded on Mendeley Data: DeDominicis, Kali (2020), "Data for: 'CStudying Attitudes Towards Vaccine Hesitance and California Law SB 277 in Online Discourse: A Dataset and Methodology'", Mendeley Data, V1, doi: 10.17632/jmpwfb9zx8.1 http://dx.doi.org/10.17632/jmpwfb9zx8.1

(*continued on next page*)

Related research article	DeDominicis, K, Buttenheim, A., Howa, A.C., Delamater, P.L., Salmon, D., Omer, S.B., & Klein, N.P. 2020. "Shouting at Each Other into the Void: A Semantic Network Analysis of Vaccine Hesitance and Support in Online Discourse Regarding California Law SB277. *Social Science & Medicine* 266. https://doi.org/10.1016/j.socscimed.2020.113216

Value of the Data

- These data represent the network of information that would be available to parents attempting to research California's 2015 vaccine mandate, SB277, or the efficacy of vaccines overall. The dataset can be used to further research and understanding of how people discuss vaccines and vaccine mandates online, and particularly illuminate possible misconceptions or misunderstandings about the motivations and priorities of the parties involved in such discourse.
- The dataset offers a baseline for understanding how online actors discuss vaccination and vaccine mandates, and how they test arguments and narratives that are intended to persuade specific audiences, both online and offline. The dataset can be used as a point of reference for other studies concerning attitudes towards or discussions of vaccine hesitance, vaccine policy, and controversial legislation more broadly.
- The dataset can be used to further analyze particular aspects of the online discourse and narrative framings of issues surrounding SB277. These data can also contribute to more general research regarding people's attitudes towards vaccines, government mandates, and the social actors sometimes known as the "anti-vaxx movement." Furthermore, this stage of analysis subjected these data primarily to manual analysis; they could benefit from computer-based sentiment and linguistic analysis.
- These data can be used by researchers interested in extending understanding of how different groups of people and different media sources present information about vaccines and vaccine mandates, and how various groups make linguistic and narrative choices about how to frame their arguments.
- The mixed methods approach used to collect and analyze these data (including a modified form of snowball sampling, linguistic frequency, semantic network, and content analysis) are more broadly helpful to researchers attempting to examine online discourse surrounding controversial practices or policies in a more general sense, particularly if they are taking a semantic network analysis approach or using other forms of linguistic analysis.

1. Data Description

The tables below reflect six "Clusters" or semantic networks of words that are related to or relevant to each other. The additional data available on Mendeley represents the raw number of times a word appeared in each attitude category; these lists were derived from 3.38 GB of raw data – all the documents scraped and uploaded into NVivo during initial data collection, including attitude coding. The raw data is ideal for additional computer-based coding and analysis for any researchers interested in online discussion of California SB277, policy mandates in general, and attitudes towards vaccination more broadly.

The final dataset comprised 2424 documents from 213 websites. It was composed of web pages, Facebook statuses and public groups, newspaper articles, blog posts, government reports, and forum discussions. There were 433 "vaccine supportive" documents (2,894,317 words), 1717 "vaccine skeptical" documents (8,055,558 words), and 274 "neutral" documents (1,602,322 words). The final list of 79 words was divided into 6 Clusters (Tables 1–6).

Table 1
Raw and analyzed data for Cluster 1; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
Health Choice	97	8.4	361	0.0045	12	4.4	12	0.0007	8	1.9	10	0.0003
Medical Choice	62	5.4	157	0.0019	8	2.9	13	0.0008	12	2.8	17	0.0006
Vaccine Choice	179	15.6	569	0.0071	6	2.2	7	0.0004	39	9.1	113	0.0039
Health Freedom	154	13.4	451	0.0056	3	1.1	3	0.0002	34	8	60	0.0021
Medical Freedom	126	11	444	0.0055	8	2.9	62	0.0039	9	2.1	10	0.0003
Vaccine Freedom	29	2.5	52	0.0006	1	0.4	1	0.0001	6	1.4	8	0.0003
Health Rights	4	0.3	10	0.0001	1	0.4	1	0.0001	1	0.2	1	0.0000
Medical Rights	12	1	37	0.0005	2	0.7	3	0.0002	1	0.2	1	0.0000
Vaccine Rights	28	2.4	64	0.0008	0	0	0	0.0000	2	0.5	2	0.0001
Vaccine Advocacy	4	0.3	4	0.0000	3	1.1	4	0.0002	14	3.3	19	0.0007
Medical Fascism	45	3.9	82	0.0010	2	0.7	3	0.0002	5	1.2	6	0.0002
Medical Tyranny	127	11.1	304	0.0038	10	3.6	13	0.0008	11	2.6	15	0.0005
Anti-vaxx	62	5.4	2496	0.031	22	8	724	0.045	73	17.1	1297	0.045

Table 2
Raw and analyzed data for Cluster 2; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
Autism	502	43.7	12,419	0.154	107	39.1	3061	0.191	252	59	4755	0.164
Child	567	49.3	39,409	0.489	260	94.9	9913	0.619	404	94.6	16,488	0.570
Parent	188	16.4	16,450	0.2	243	88.7	3775	0.24	369	86.4	7008	0.24
Measles	442	38.5	18,818	0.234	217	79.2	4799	0.300	321	75.2	8839	0.305
Science	540	47	8570	0.106	167	60.9	2118	0.132	323	75.6	6137	0.212
Education	709	61.7	8460	0.105	182	66.4	1307	0.082	278	65.1	2224	0.077
School	87	7.6	3303	0.041	23	8.4	2452	0.153	61	14.3	4200	0.145
PBE	82	7.1	3859	0.048	48	17.5	2760	0.172	67	15.7	2651	0.092
Religious Exemption	35	3	433	0.005	16	5.8	34	0.002	24	5.6	50	0.002
Disneyland	70	6.1	1074	0.013	54	19.7	236	0.015	59	13.8	372	0.013
Civil Rights	108	9.4	1268	0.016	14	5.1	76	0.005	43	10.1	298	0.010
Human Rights	193	16.8	2456	0.030	24	8.8	443	0.028	42	9.8	615	0.021
Children's Rights	33	2.9	74	0.001	8	2.9	14	0.001	20	4.7	30	0.001
Parents' Rights	310	27	8421	0.105	73	26.6	2552	0.159	103	24.1	4356	0.151

2. Experimental Design, Materials and Methods

We analyzed documents published between June 2014, six months before the Disneyland measles outbreak [2], and June 2017, the end of the first school year after SB277 was implemented [3]. That broad date range was chosen to enable future analysis and comparison of different eras of response to SB277: Prior to the Disneyland outbreak, after the outbreak but prior to the proposal of a policy solution, during the discussion and passage of SB277, between the passage of the bill and its implementation, and during the first year of implementation. This first analytical paper, however, deals with the data as a whole, to provide a baseline for comparison. The data represent information that would have been available to parents at the time, and includes only naturally-occurring discourse surrounding SB277, rather than replies to researchers' inquiry. Only text data were analyzed; graphics and YouTube videos were excluded.

Table 3
Raw and analyzed data for Cluster 3; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
Autism	502	43.7	12,419	0.154	107	39.1	3061	0.191	252	59	4755	0.164
ADHD	102	8.9	1821	0.023	15	5.5	60	0.004	39	9.1	111	0.004
Antibodies	126	11	1194	0.015	32	11.7	343	0.021	69	16.2	43	0.001
HPV	194	16.9	1547	0.019	32	11.7	470	0.029	124	29	860	0.03
Measles	442	38.5	18,818	0.234	217	79.2	4799	0.3	321	75.2	8839	0.305
Pertussis	196	17.1	2681	0.033	70	25.5	793	0.049	161	37.7	1789	0.062
Smallpox	129	11.2	899	0.011	43	15.7	338	0.021	101	23.7	624	0.022
Virus	277	24.1	5591	0.069	75	27.4	1314	0.082	158	37	1523	0.053
Disease	640	55.7	17,366	0.216	215	78	4537	0.283	117	27	8167	0.282
Aluminum	247	21.5	3228	0.04	45	16.4	573	0.036	71	16.6	750	0.026
Mercury	307	26.7	2414	0.03	65	23.7	2890	0.18	96	22.5	1039	0.036
Thimerosal	136	11.8	1198	0.015	36	13.1	471	0.029	78	18.3	545	0.019

Table 4
Raw and analyzed data for Cluster 4; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
America	639	55.6	6271	0.08	142	51.8	876	0.05	245	57.4	1609	0.06
California	303	26.4	11,192	0.14	77	28.1	2895	0.18	116	27.2	1718	0.06
Government	729	63.4	9343	0.12	125	45.6	1274	0.08	228	53.4	1707	0.06
Big Government	63	5.5	794	0.01	8	2.9	430	0.03	11	2.6	272	0.01
CDC	568	49.4	16,961	0.21	103	37.6	6457	0.4	212	49.6	10,422	0.36
FDA	240	20.9	3077	0.04	46	16.8	952	0.06	115	26.9	1396	0.05
Constitutionality	424	36.9	2437	0.03	84	30.7	364	0.02	139	32.6	517	0.02
Legal	465	40.5	2645	0.03	81	29.6	430	0.03	177	41.5	686	0.02
Illegal	187	16.3	909	0.01	28	10.2	89	0.01	71	16.6	186	0.01
Big Pharma	521	45.3	4411	0.055	62	22.6	880	0.055	186	43.6	1207	0.042
Pharmaceutical Industry	354	30.8	7397	0.092	52	19	1918	0.120	102	23.9	2402	0.083
Fraud	81	7	1051	0.013	16	5.8	115	0.007	29	6.8	73	0.003

2.1. Sampling

Data collection began with Google searches for “SB277,” “California SB277,” “California vaccine bill,” and “Senate Bill 277.” This yielded 13 websites that belonged to organizations associated with SB277 or general vaccine activism [1]. We developed a modified form of “snowball sampling” [4] to expand the data beyond those 13 sites. In traditional snowball sampling, researchers ask informants to recommend additional contacts who might be able to contribute to a study, thus partially reflecting the organic structure and development of social networks.

Our modified digital snowball sampling method was designed to reflect the digital networks of recommendations and linked webpages that individuals can access and participate in Fig. 1. Google site search was used to identify individual pages within each of the 13 starting websites that mentioned SB277 specifically, which excluded irrelevant pages (e.g. contact information, donation appeals). Those pages were “scraped” and uploaded to the NVivo Suite [5]. Then, every link on each relevant page was opened, excluding irrelevant items (e.g. advertisements). Some of the new pages revealed through these links were on “outside” websites not included in the original 13. The Google site search process was repeated for each new website, revealing not just the

Table 5
Raw and analyzed data for Cluster 5; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
Nazi	50	4.4	398	0.0049	8	2.9	28	0.0017	24	5.6	116	0.004
Hitler	29	2.5	125	0.0016	6	2.2	17	0.0011	18	4.2	54	0.0019
Fascism	53	4.6	189	0.0023	8	2.9	24	0.0015	13	3	63	0.0022
Anti- Semitism	6	0.5	27	0.0003	1	0.4	6	0.0004	18	4.2	74	0.0026
Jewish	29	2.5	324	0.004	5	1.8	13	0.0008	27	6.3	168	0.0058
Holocaust	21	1.8	308	0.0038	9	3.3	23	0.0014	25	5.9	185	0.0064
Nuremberg Code	35	3	125	0.0016	4	1.5	9	0.0006	4	0.9	27	0.0009
Christian	27	2.3	206	0.0026	8	2.9	11	0.0007	25	5.9	81	0.0028
African- American	35	3	936	0.0116	6	2.2	48	0.003	9	2.1	72	0.0025
Nation of Islam	11	1	42	0.0005	5	1.8	7	0.0004	9	2.1	22	0.0008
Muslim	27	2.3	166	0.0021	8	2.9	15	0.0009	18	4.2	54	0.0019
Tuskegee	14	1.2	142	0.0018	1	0.4	2	0.0001	7	1.6	9	0.0003
Racism	18	1.6	107	0.0013	5	1.8	25	0.0016	11	2.6	42	0.0015

Table 6
Raw and analyzed data for Cluster 6; D% refers to the total number of documents in which that phrase appears within an attitude category relative to the total number of documents in that attitude category. R% refers to the total number of times a word or phrase appears in an attitude category in total, relative to the total number of words in that attitude category.

	Skeptical				Neutral				Supportive			
	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%	Docs.	D%	Refs.	R%
Sen. Pan	392	34.1	2297	0.029	140	51.1	373	0.023	119	27.9	246	0.008
Gov. Brown	248	21.6	1489	0.018	99	36.1	1570	0.098	111	26	305	0.011
Donald Trump	252	21.9	1306	0.016	73	26.6	173	0.011	115	26.9	1083	0.037
Dorit Reiss	20	1.7	167	0.002	29	10.6	449	0.028	132	30.9	968	0.033
Sen. Allen	80	7	146	0.002	58	21.2	65	0.004	52	12.2	73	0.003
Dr. Sears	95	8.3	9515	0.118	22	8	892	0.056	102	23.9	1148	0.040
Sen. Kennedy	97	8.4	760	0.009	27	9.9	352	0.022	66	15.5	787	0.027
Jenny McCarthy	28	2.4	250	0.003	26	9.5	56	0.003	86	20.1	208	0.007
CDC Whistleblower	151	13.1	2037	0.025	22	8	854	0.053	43	10.1	1059	0.037
Dr. Wakefield	55	4.8	1027	0.013	27	9.9	207	0.013	63	14.8	318	0.011
Hillary Clinton	115	10	924	0.011	16	5.8	51	0.003	33	7.7	83	0.003
Dr. Offit	41	3.6	700	0.009	12	4.4	43	0.003	38	8.9	111	0.004
Dr. Tenpenny	59	5.1	232	0.003	7	2.6	23	0.001	16	3.7	39	0.001
Jim Carrey	10	0.9	52	0.001	6	2.2	13	0.001	25	5.9	398	0.014
Dr. Obukhanych	26	2.3	335	0.004	5	1.8	228	0.014	8	1.9	505	0.017
Jenna Elfman	14	1.2	233	0.003	6	2.2	27	0.002	4	0.9	11	0.000

pages the previous documents had linked to, but any additional pages on the new website that mentioned SB277 or our other search terms. For example, if a Facebook group linked to articles on Infowars and the *Sacramento Bee* newspaper website, those specific documents were added to the dataset, and so were any other articles that Google identified as mentioning "SB277" on both Infowars.com or SacBee.com. The links on all new pages were then opened, and the process repeated.

This approach was intended to retroactively approximate aspects of parents' experiences researching SB277. That is, the dataset reflected a web of links constructed to include all possible webpages a parent might have arrived at by following initial Google searches that resembled ours, or from a Google search that yielded any of the subsequent pages instead. Thus, the experience of a parent who began on a parenting blog instead of one of the initial website pages

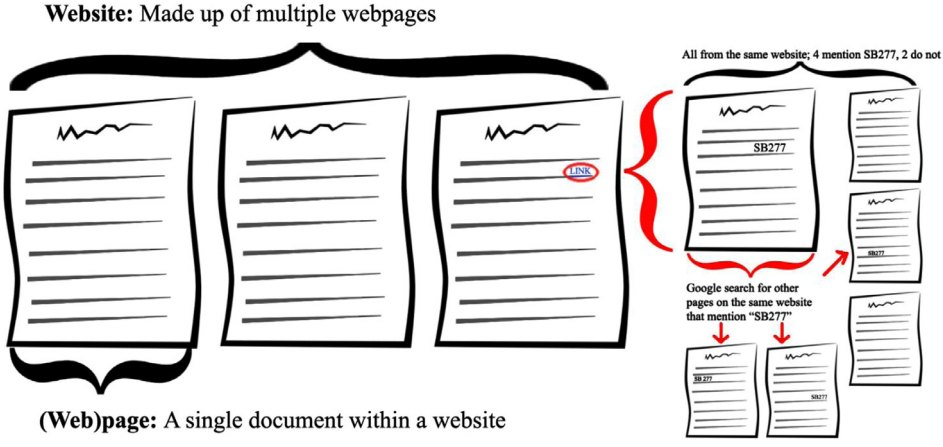

Fig. 1. Illustration of snowball sampling.

represented by our search would still be included in our dataset, with the understanding that they would have access to the same network of documents, but starting at a different nexus.

2.2. Analysis and coding

Each unique page was considered a "document." Once uploaded to NVivo, documents were manually coded into qualitative categories according to the following attitudes: "vaccine supportive," "vaccine skeptical," or "neutral." For this initial article, manual coding was preferred because, even with the exclusion of video and graphic data, the documents produced through snowball sampling were highly representative of the multimodal, heterogeneous nature of digital texts; as such, the data and analysis in these paired papers establish a manual baseline that can enable future machine-based coding and analysis techniques. The supportive and skeptical attitude categories were *a priori* components of the analysis, because this research always intended to compare those two perspectives, but the neutral category emerged in analysis in order to represent texts such as news reports that made efforts not to espouse either position. Supportive documents were defined as those that actively asserted the medical benefits of vaccines and the scientific basis thereof, or included documents supportive of SB277 as a policy. Skeptical documents proclaimed the dangers of vaccines or vaccine mandates. Neutral documents did not attempt to influence readers towards either position, and expressed no implicit or explicit opinion about the efficacy of the bill or vaccination itself.

Initial coding decisions were guided by the explicit attitudes of host websites: most authors stated their vaccine supportive or hesitant views up front, in the document or on the broader website, and most documents were coded according to that stated position. These assertions were taken at face value for three reasons. First, the majority of skeptical and supportive websites were consistent platforms: they rarely hosted documents with dissenting points of view, but rather sought to present consistent, convincing narratives. In order to catch rare exceptions to this practice, researchers manually checked the title of each document against the stated attitude of its host website and made corrections as necessary; for documents with ambiguous titles, light content analysis informed the ultimate decision. Second, no discrepancies in attitude coding were uncovered during the detailed content analysis stage, when researchers had the opportunity to check the subtext of each document against its attitude code; this suggests high overall reliability in initial coding. Third, while it was possible that individual authors or websites might misrepresent their positions, this methodology was partly intended to reflect

the experiences of information-gathering parents, who would have been no more equipped to discern insincerity than the researchers (see [6]).

Neutral documents were harder to code. As noted above, this category emerged during analysis in order to include documents that took no explicit stance regarding SB277 or vaccines overall. Neutral documents primarily consist of government information sheets about the basic facts of SB277 and journalistic publications that tried to represent citizens from both camps equally without rendering judgement on their opinions. Within these strictures, documents with titles that suggested particular stances, publications or websites that seemed to consistently represent only one position, and documents that were denoted as editorials or similar were subjected to light content analysis before final coding. Detailed content analysis only revealed two errors in neutral coding (out of 274 total documents), which suggests high overall reliability.

Comment sections, in which potentially infinite numbers of individuals could respond to the original text and to other commenters, were too complex and lengthy for manual coding to be feasible. Comment sections are frequent sites of conflict, particularly in digital spaces that reflect contentious issues, such as vaccine efficacy or policy mandates: thus, each addition to an exchange would have to be coded individually. Machine-based coding and sentiment analysis can render more detailed categorizations of comment data, and this could be a fruitful avenue for future analysis, but the manual baseline provided by this stage of analysis are an important first step for such techniques.

For this stage of analysis, comment sections were left *in situ* to preserve context; when comments were embedded at the bottom of a page, the main text and comments were considered a single document. When comments were accessible through a link to a separate page, they were considered two documents. This had certain implications for attitude coding. First, embedded comments were necessarily coded according to the attitude of the original text because they were not treated as discrete documents. In order to preserve continuity between embedded and discrete comments, we also coded the latter according to the attitude of the original text. More importantly, coding comment sections according to the attitude of the original text preserved the context in which readers, including concerned parents, might encounter these views; reading vaccine skeptical narratives in a neutral digital space (e.g., attached to a newspaper article) carries a different weight than finding them attached to a vaccine skeptical blog.

This coding decision meant that comments sometimes represented views that conflicted with the attitude they had been coded as holding. This issue was addressed by separately coding relevant documents or sections of documents as "comments." We performed two analyses, once including and once excluding comment sections, which allowed us to examine how the two differed. In this data, only Cluster 4 showed a notable disparity between the two analyses, but a more detailed examination of how comment sections alter the data could be an interesting topic for future inquiry (see Table 4). Neutral documents could particularly benefit from this, because the technical and philosophical design of comment sections is about enabling readers to express opinions and persuade each other – which is at odds with our definition of neutral documents. Furthermore, about half of the documents classified as "neutral" are or include comment sections; this is because national news platforms have a larger audience than even most popular bloggers, Facebook groups, etc., so they had a larger source of potential commenters.

NVivo was used to identify the 500 most common terms in each attitude category separately and for the dataset as a whole. This included phrases of up to three words. The NVivo "stemmed" function combined coding for words with the same root, while manually-determined related words were also coded together. Thus, the individual words in our data actually represent a semantic network of similar words; for example, the word "child" also represents "children," "kid," "childhood," etc. Filler words were eliminated, as were certain overly-generic words; for example, "vaccine," "bill" and "people" merely demonstrate that these were discussions of vaccine policy, supported or opposed by groups of people. After these eliminations, the top ten words and top ten phrases in each attitude category were selected for further analysis. Terms that appeared in the top twenty of at least two categories were also included.

For each significant term, in-depth content analysis was applied to the five documents with the highest concentrations of that word. This was partly because word frequency was a key

aspect of how we chose data for analysis, so using another system to choose documents for content analysis risked obscuring the texts that had caused particular words to become notable. Primarily, however, this decision was efficiency-based: texts with high concentrations of a word provided more examples of how that word was used, how it was framed, and helped develop a qualitative idea of the semantic and narrative networks that surrounded it.

Deeper, qualitative readings were undertaken using a documentary and textual analysis framework, with particular attention paid to how each term was used in context, and what issues and events they were most relevant to. This reading suggested that a semantic network approach (SNA) would improve our analysis: simply comparing high frequency words to each other would be less fruitful than using SNA to identify "emergent clusters of potential meaning by analyzing relations between words" [7]. For example, "parents' rights" and "[Dr. Robert] Sears" were both prominent terms, but it was less useful to compare them than to contrast the former to "civil rights," "human rights," and "children's rights."

Semantic networks were constructed by drawing additional terms from the original lists of 500 words; these selections were guided by the NVivo [5] "word tree" SNA function and the context provided by initial qualitative content analysis. Most words were still in the top 20–50 for each attitude category, and the words that did not meet that frequency criteria were chosen for being part of the same semantic network as the high frequency words. In effect, the relative unpopularity of words that were linguistically or conceptually similar to high frequency words were included when they illuminated interesting aspects of the narrative choices and social framings espoused by the authors. These networks were divided into "clusters" of semantic networks relevant to the initial set of high frequency words.

There were substantially more skeptical documents than other types, which impeded direct comparisons between the three categories. Thus, percentages were calculated. This was done in two ways: first, the number of documents each keyword appears in, in each attitude category, out of the total number of documents in that category (D%). This demonstrated how widespread or general concern with that topic was among people with that attitude. For some clusters, we also calculated the number of times a word appeared in total per attitude category, out of the total number of times the word was used in that category (R%). We calculated R% in all cases [see Tables], but this paper only includes charts of R% for clusters where there was a disparity between how often a word appeared in total versus how many documents it appeared in. For example, in Cluster 3 the word "mercury" appears in 23.7% of neutral documents, which is within 3% of the other categories – but it is 0.18% of all words in the neutral category, which is 5 times more frequent than in skeptical or supportive sources (see Table 3).

Ethics Statement

This dataset relies heavily on social media data, including Facebook posts, blog posts, and newspaper comments. As such, hundreds of ordinary citizens' legal names are present in the full dataset, in whole or in part, with no reliable (or practical) way to obtain informed consent for their use. Furthermore, even people who use "screen names" or other anonymous handles online may have a long-term attachment to a particular digital identity; thus, including even obviously fake names in a dataset intended for public use is not necessarily much more ethical than including their legal identities [8].

All of these exchanges took place in public forums: they were, as the methodology demonstrates, obtainable through ordinary Google searches, and did not require any additional credentials to access them. While some digital scholars argue that the public nature of such discourse constitutes consent, or at least ethical justification for using such documents in academic research, there is no broad consensus as to whether participation in a public online forum necessarily indicates that participants understand that those forums *are* public or when and how researchers should be able to use their data [9,10]. More importantly, it is impossible to anonymize the full dataset in a way that protects the ordinary people who are represented in it. [8-10]. For this reason, the final dataset shared in this *Data in Brief* submission is the full complement of

word lists used in our analysis and not the complete set of documents collected in the initial research. Researchers who are interested in utilizing the full set of documents should contact the corresponding author to discuss ethical precautions.

Finally, while the dataset associated with this article began as NVivo-generated [5] lists of the 1000 most common words and the 750 most common stemmed words in all attitude categories (as well as the total), the totals on some lists are lower than that due to omitting recognizable names and screen names. The only exceptions are for public figures who actively involved themselves in this debate on a state or national level, including politicians, doctors who consult on government policy or speak to national news organizations, and celebrities. Common names, such as “Mike” or “Elizabeth” were left in as well, since they do not endanger people’s identities out of context.

This is all in compliance with the original IRB clearance of the study, which presumed that anonymization protocols would ensure that no individuals on social media would be identified or subjected to undue scrutiny through this dataset or associated analysis. It is also in compliance with all regulations governing all relevant social media platforms.

CRediT Author Statement

Kali DeDominicis: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Visualization, Roles/Writing – original draft, Writing – review & editing; **Alison M. Buttenheim:** Writing – Review & Editing, Validation; **Amanda C. Howa:** Project Administration, Validation, Writing – Review & Editing; **Paul L. Delamater:** Writing – Review & Editing; **Daniel Salmon:** Writing – Review & Editing; **Nicola P. Klein:** Writing – Review & Editing; **Saad B. Omer:** Writing – Review & Editing; Funding acquisition.

Declaration of Competing Interest

Nicola P. Klein has received research support from GlaxoSmithKline, Pfizer, Merck, Sanofi Pasteur and Protein Science (now Sanofi Pasteur).

Acknowledgments

This work was supported in part by the National Institutes of Health R01AI125405.

References

[1] K. DeDominicis, A. Buttenheim, A.C. Howa, P.L. Delamater, D. Salmon, S.B. Omer, N.P. Klein, Shouting at each other into the void: a semantic network analysis of vaccine hesitance and support in online discourse regarding California Law SB277, Soc. Sci. Med. 266 (2020), doi:10.1016/j.socscimed.2020.113216.
[2] M. Majumder, E. Cohn, S. Mekaru, J. Huston, J. Brownstein, Substandard vaccination compliance and the 2015 measles outbreak, JAMA Pediatr. 169 (5) (2015) 494–495, doi:10.1001/jamapediatrics.2015.0384.
[3] California Legislative Information, SB-277 Public health: Vaccinations. https://leginfo.legislature.ca.gov/faces/billHistoryClient.xhtml?bill_id=201520160SB277, 2016. Accessed March 2, 2020.
[4] F. Baltar, I. Brunet, Social research 2.0: virtual snowball sampling method using Facebook, Internet Res. 22 (1) (2012) 57–74, doi:10.1108/10662241211199960.
[5] Q.I.P. Ltd., Nvivo, Victoria, 2014.
[6] T. Boellstorff, Coming of Age in Second Life: An Anthropologist Explores the Virtually Human, Princeton University Press, NJ, 2008.
[7] J.B. Ruiz, G.A. Barnett, Exploring the presentation of HPV information online: a semantic network analysis of websites, Vaccine 35 (2015) 3354–3359, doi:10.1016/j.vaccine.2015.05.017.
[8] C. Paris, N. Colineau, S. Nepal, S.K. Bista, G. Beschorner, Ethical considerations in an online community: the balancing act, Eth. Inf. Technol. 15 (2013) 301–316.
[9] A. Franzke, A. Bechmann, M. Zimmer, C.M. Ess, Internet Research: Ethical Guidelines 3.0 (2020). https://aoir.org/reports/ethics3.pdf.
[10] K. Orton-Johnson, Ethics in online research; evaluating the ESRC framework for research ethics categorisation of risk, Soc. Res. Online 15 (4) (2010) Article 13 http://www.socresonline.org.uk/15/4/13/13.pdf.