Title: Methods supplement appendix for “Google allows advertisers to target the sensitive informational queries of cancer patients”
Authors: Marco Zenone (1), Alessandro Marcon (2), Nora Kenworthy (3), May van Schalkwyk (1), Timothy Caulfield (2), Greg Hartwell (1), Nason Maani (4)
Date: December 23rd, 2024
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix: Methods supplement

Data collection

Data pull 1: The image below demonstrates the “domain overview” dashboard which was searched for each alternative cancer clinic. The chart in the green box was exported to CSV for each clinic and aggregated with the other clinics to understand cumulative website organic and paid traffic, as well as paid traffic cost. The clinic searched in this example is anonymized.

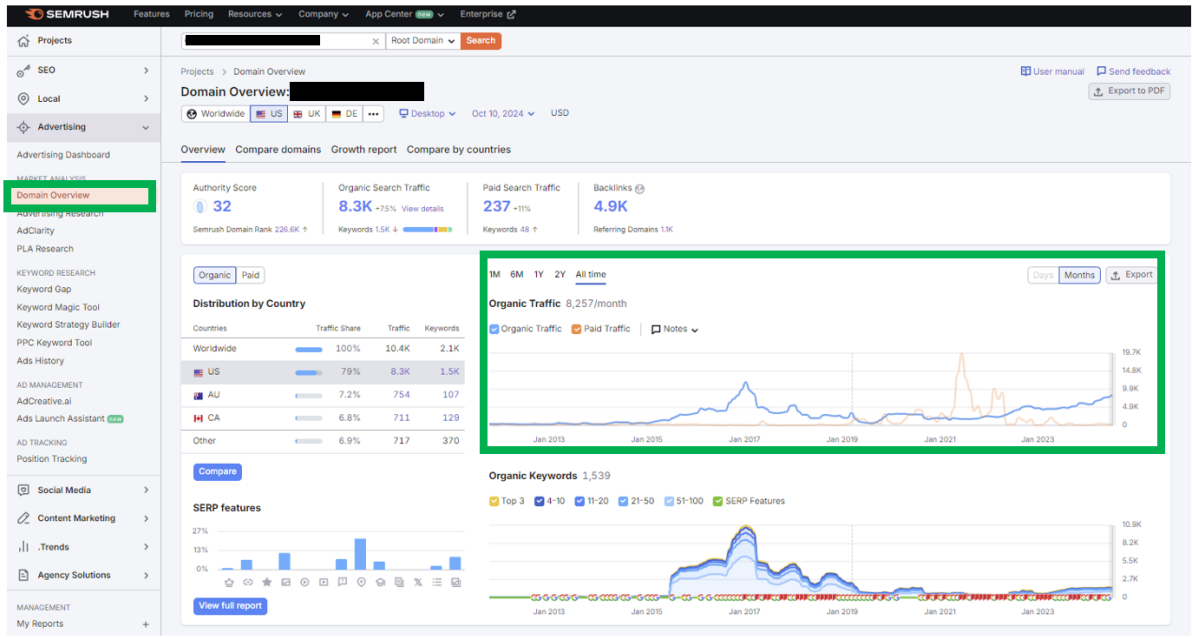

Figure A1. Collection of alternative cancer clinic organic traffic, paid traffic, and paid traffic cost.

Data pull 2: The image below demonstrates the “advertising research” dashboard which was searched for each alternative cancer clinic by each month of 2023. The bottom part of the image shows the exportable keywords, which were exported to CSV. There were other data that was possible to collect but we only were interested in the keywords used for the purpose of this analysis. After exporting the keywords used by each alternative cancer clinic by month in 2023, we removed the duplicates, identifying 20,035 unique keywords. The clinic searched in this example is anonymized.

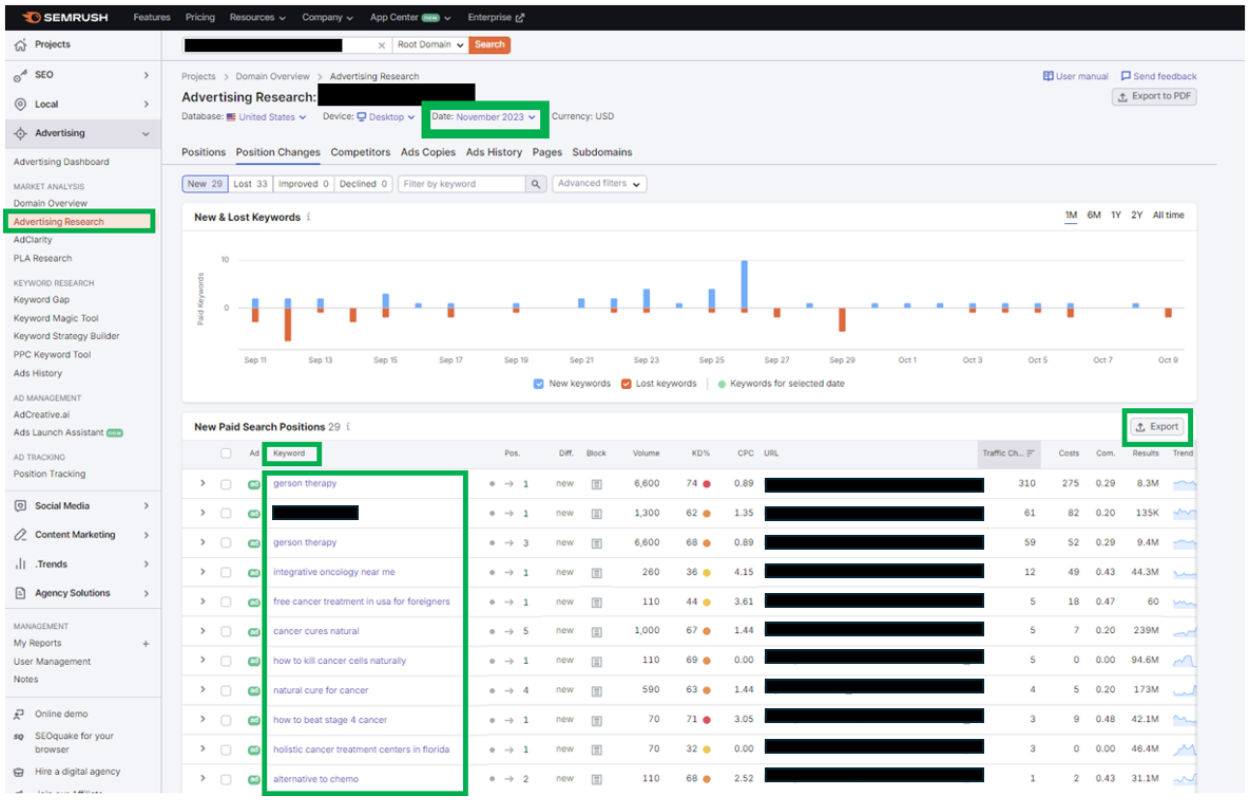

Figure A2. Collection of alternative cancer clinic paid Google search ad keywords.

Analysis

MZ independently reviewed half of the keywords to develop an inductive coding frame. Each keyword was considered a unique unit for coding. MZ found, iteratively reworked, refined, and defined themes. MZ repeated this process until no other themes were found or characterized within existing themes. Throughout this process, MZ coded 15,000 of the keywords and stopped upon reaching saturation of themes. AM completed two audits of a generalizable random sample of the keyword data (n = 400) of the code frame to verify saturation of the thematic categories and to test the defining characteristics of each. The first audit led to minor changes brought about through discussions between MZ and AM. The second audit tested the changes anew to confirm the presence of new categories, and to verify the completeness of defining characteristics. The second audit led to the identification of no new themes across the keywords.

Limitations

The limitation of the study is its data source. Semrush uses several methods to build its metrics but does not have access to the specific Google accounts of alternative cancer clinics. Therefore, while the data can provide useful insights, there are limitations to its accuracy. Semrush data collection is informed by partnerships between the company and clickstream data providers and uses its neural network algorithm to estimate traffic based on statistical sampling. Ads data are retrieved from databases tracking Google ads and estimating spend and reach. The data source and methods employed in the study are the most appropriate to answer our research questions. At present, there are few options to study how paid Google ads are targeted to cancer patients. Google provides in its Transparency Center a partial repository of ads which provide no information on ad targeting besides country and no engagement or spend metrics.