New York University Abu Dhabi

Applied Data Science

Utilizing Machine Learning Models to Effectively Predict Key Prevalence Rates in Public Health

Group 1:

Eva Bi, Lihan Feng, Sean Shan, Vickie Wang

Word Count: 2474

Abstract

This study explores the innovative use of Google search keywords as a predictive tool for estimating the prevalence rates of obesity and exercise behaviors. Leveraging various machine learning models, including Random Forest, Lasso, Ridge, and Polynomial, and we developed a framework to analyze the relationship between search query frequencies and epidemiological trends (i.e., obesity and exercise rates). Our methodology involved extensive training and evaluation of these models to ensure accurate predictions. The performance of each model was rigorously assessed, highlighting the strengths and limitations inherent in different approaches. The results demonstrate a promising alignment between search keyword trends and actual prevalence rates, suggesting that Google searches can serve as a cost-effective and timely alternative to traditional demographic surveys. This research contributes to the field by offering an accessible method for health surveillance and public health policy planning, harnessing the ubiquity and immediacy of internet search data. Our findings underscore the potential of machine learning in public health analytics, opening avenues for real-time monitoring and proactive health management strategies.

Introduction

In the contemporary landscape of public health in the United States, the interplay between obesity and physical exercise emerges as a critical area of concern. The prevalence of obesity, a leading risk factor for numerous chronic diseases, has been steadily escalating, underscoring the urgency for effective public health strategies. Obesity is linked to several serious health conditions, including heart disease, type 2 diabetes, stroke, certain cancers, severe outcomes from COVID-19, and poor mental health. It is also important to note that adults with obesity are at increased risk for these conditions and often face weight stigma, which can negatively impact their healthcare experiences and outcomes. Similarly, the importance of regular exercise in promoting overall health and well-being is widely recognized, yet its practical application often remains underutilized. Research consistently highlights the crucial role of physical exercise in enhancing physical fitness, reducing the risk of various chronic diseases, and improving mental health. The integration of regular physical activity into daily life is seen as a cornerstone for a healthy lifestyle, yet the challenge remains in effectively translating this knowledge into consistent, practical actions across diverse populations.

Existing papers that focus on evaluating correlations and monitoring obesity and exercise rates often involve direct measurement or observation of physical activity levels, fitness, and dietary behaviors in specific populations. For instance, some studies focus on directly assessing body fat and obesity using various techniques, including measuring body mass index (BMI), percentage body fat, and other anthropometric measures to evaluate obesity levels in different populations. By using such objective measures, researchers can gain a clearer understanding of how physical activity influences obesity and vice versa. While these traditional methods of examining physical activity and obesity can provide accurate and direct data, they come with certain disadvantages. One limitation is the resource-intensive nature of these methods, often requiring extensive manpower, time, and financial investment for data collection. Furthermore, these methods may not capture the rapidly changing patterns of public behavior and attitudes, especially on a large scale. Additionally, direct measurement approaches can be intrusive and may not be feasible for large population studies due to logistical challenges and potential privacy concerns.

In contrast, this study adopts search trends as the primary methodological approach, leveraging its unique advantages. The wide reach and accessibility of search data allow us to capture a broad spectrum of public engagement across diverse demographics, offering an expansive and comprehensive perspective on health-related behaviors and interests. This approach is particularly cost-effective, allowing for extensive data collection without incurring the significant expenses typically associated with traditional research methods. Furthermore, search trends serve as a valuable indicator of public awareness and interest, especially in health domains such as exercise and obesity, thus providing crucial insights into prevailing societal attitudes and behaviors. The predictive capability of search trends analysis is another advantage, enhancing the ability to forecast emerging public health trends and facilitating proactive measures in health policy and management.

These combined factors make search trends analysis a robust and efficient tool for investigating and understanding complex health-related phenomena, addressing the limitations of direct measurement methods while providing broad and insightful data.

By leveraging readily available data sources, this research seeks to answer the question:
Can search trends related to obesity and exercise effectively predict the actual rates of obesity and exercise participation in the United States over a specified period (2005-2018)?
This inquiry not only addresses a crucial public health issue but also proposes a methodological shift towards more efficient and accessible data utilization in health trend analysis.

Methodology and Results

Data Collection and Preprocessing

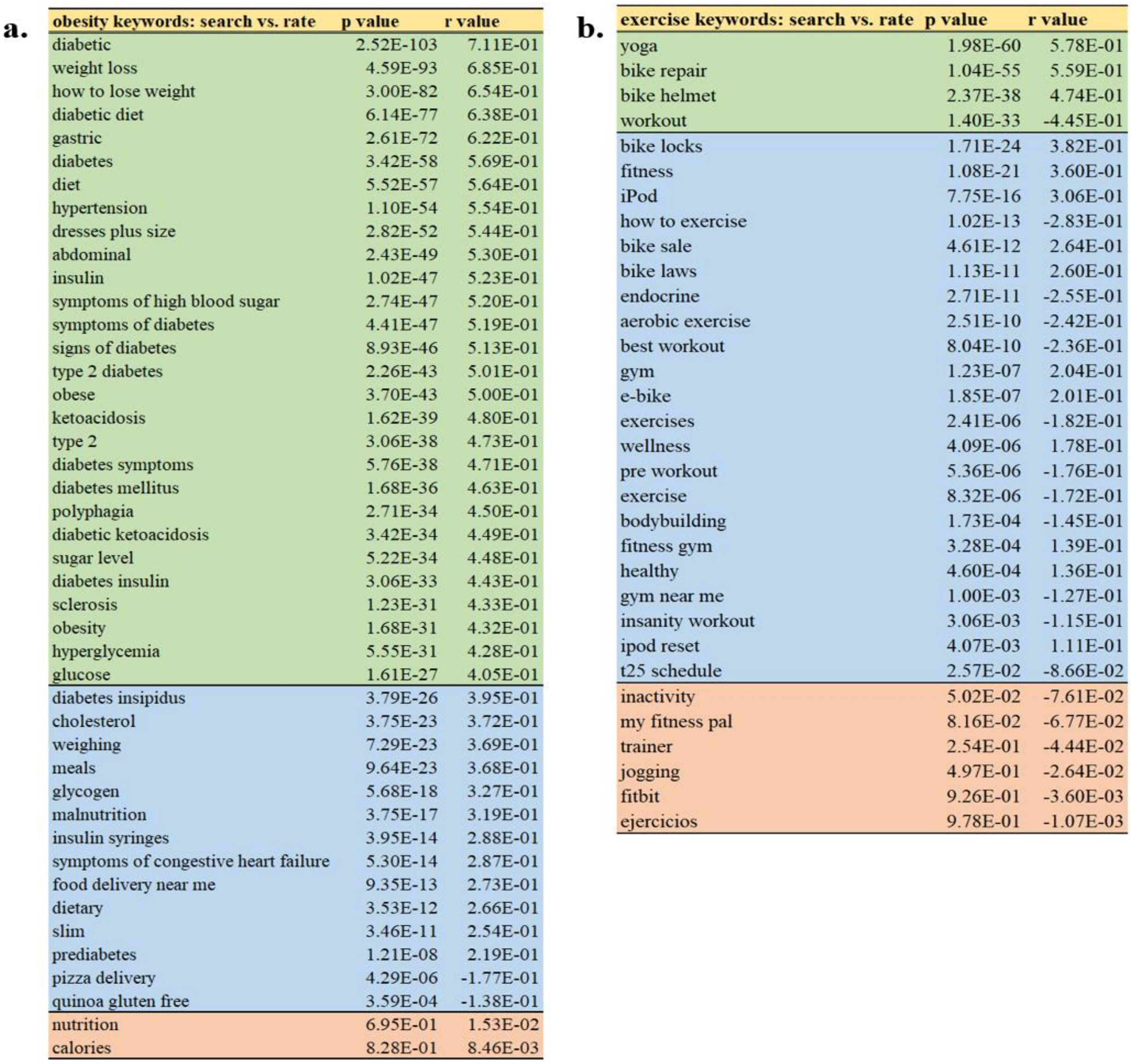
In this study, we leveraged the Health Dataset by Memon, Razak, and Weber from the Journal of Medical Science Research, encompassing U.S. state-level Obesity and Exercise Prevalence Rates from 2004 to 2018, alongside 81 Google search keywords. Rigorous data preprocessing was essential for machine learning model training, involving dataset cleaning, inconsistency rectification, and appropriate formatting.

Prevalence Rates dataset preprocessing involved eliminating noisy data and focusing on crucial columns: years, states, prevalence rates, stratifications (overall, age-based, gender-based), and types (obesity, exercise). Duplicate or outlier checks showed no significant issues. NaN values, under 0.3% of the dataset, were removed. The Google search dataset was limited to overall values, with detailed stratification noted as a limitation. The datasets were categorized into obesity and exercise, resulting in two distinct, preprocessed datasets. In processing the Google search keywords dataset, we evaluated and excluded five irrelevant keywords ('NIH', 'apnea', 'unhealthy', 'visceral', 'icd 10 codes'). The remaining keywords were categorized into obesity- (44 keywords) and exercise-related (32 keywords), forming two separate datasets, as detailed in the Appendix. We then integrated the Google search keywords with the prevalence rates datasets. This integration ensured comprehensive geographic coverage, including all 50 U.S. states and Washington D.C., from 2004 to 2018. The absence of data for Hawaii in 2004 led to the exclusion of 2004 data, maintaining robustness from 2005 to 2017 for model training and validation, using the keywords as predictive features for obesity and exercise prevalence rates, with testing up to 2018.

This meticulously preprocessed and integrated dataset was optimized for machine learning applications. The subsequent steps in our study involved feature engineering and the training of machine learning models, utilizing this unified dataset. The detailed attention paid to preprocessing assured the quality and reliability of the data, laying a solid foundation for accurate and insightful machine learning analysis.

Machine Learning Feature Engineering

In the feature engineering phase of our machine learning study, we conducted a detailed correlation analysis to discern the relationship between Google search keywords and prevalence rates of obesity and exercise. This process entailed calculating Pearson correlation coefficients for each feature-target pair. A critical aspect of our methodology was the adoption of a stringent p-value threshold of less than 0.05 for feature retention. This criterion, a standard in scientific research, was instrumental in reducing the probability of false positives, thereby bolstering the credibility and relevance of our model. Our focus on statistically significant correlations played a vital role in enhancing the predictive accuracy of our models, ensuring that the analysis was firmly anchored in data of statistical significance.



a.

obesity keywords: search vs. rate	p value	r value
diabetic	2.52E-103	7.11E-01
weight loss	4.59E-93	6.85E-01
how to lose weight	3.00E-82	6.54E-01
diabetic diet	6.14E-77	6.38E-01
gastric	2.61E-72	6.22E-01
diabetes	3.42E-58	5.69E-01
diet	5.52E-57	5.64E-01
hypertension	1.10E-54	5.54E-01
dresses plus size	2.82E-52	5.44E-01
abdominal	2.43E-49	5.30E-01
insulin	1.02E-47	5.23E-01
symptoms of high blood sugar	2.74E-47	5.20E-01
symptoms of diabetes	4.41E-47	5.19E-01
signs of diabetes	8.93E-46	5.13E-01
type 2 diabetes	2.26E-43	5.01E-01
obese	3.70E-43	5.00E-01
ketoacidosis	1.62E-39	4.80E-01
type 2	3.06E-38	4.73E-01
diabetes symptoms	5.76E-38	4.71E-01
diabetes mellitus	1.68E-36	4.63E-01
polyphagia	2.71E-34	4.50E-01
diabetic ketoacidosis	3.42E-34	4.49E-01
sugar level	5.22E-34	4.48E-01
diabetes insulin	3.06E-33	4.43E-01
sclerosis	1.23E-31	4.33E-01
obesity	1.68E-31	4.32E-01
hyperglycemia	5.55E-31	4.28E-01
glucose	1.61E-27	4.05E-01
diabetes insipidus	3.79E-26	3.95E-01
cholesterol	3.75E-23	3.72E-01
weighing	7.29E-23	3.69E-01
meals	9.64E-23	3.68E-01
glycogen	5.68E-18	3.27E-01
malnutrition	3.75E-17	3.19E-01
insulin syringes	3.95E-14	2.88E-01
symptoms of congestive heart failure	5.30E-14	2.87E-01
food delivery near me	9.35E-13	2.73E-01
dietary	3.53E-12	2.66E-01
slim	3.46E-11	2.54E-01
prediabetes	1.21E-08	2.19E-01
pizza delivery	4.29E-06	-1.77E-01
quinoa gluten free	3.59E-04	-1.38E-01
nutrition	6.95E-01	1.53E-02
calories	8.28E-01	8.46E-03

b.

exercise keywords: search vs. rate	p value	r value
yoga	1.98E-60	5.78E-01
bike repair	1.04E-55	5.59E-01
bike helmet	2.37E-38	4.74E-01
workout	1.40E-33	-4.45E-01
bike locks	1.71E-24	3.82E-01
fitness	1.08E-21	3.60E-01
iPod	7.75E-16	3.06E-01
how to exercise	1.02E-13	-2.83E-01
bike sale	4.61E-12	2.64E-01
bike laws	1.13E-11	2.60E-01
endocrine	2.71E-11	-2.55E-01
aerobic exercise	2.51E-10	-2.42E-01
best workout	8.04E-10	-2.36E-01
gym	1.23E-07	2.04E-01
e-bike	1.85E-07	2.01E-01
exercises	2.41E-06	-1.82E-01
wellness	4.09E-06	1.78E-01
pre workout	5.36E-06	-1.76E-01
exercise	8.32E-06	-1.72E-01
bodybuilding	1.73E-04	-1.45E-01
fitness gym	3.28E-04	1.39E-01
healthy	4.60E-04	1.36E-01
gym near me	1.00E-03	-1.27E-01
insanity workout	3.06E-03	-1.15E-01
ipod reset	4.07E-03	1.11E-01
t25 schedule	2.57E-02	-8.66E-02
inactivity	5.02E-02	-7.61E-02
my fitness pal	8.16E-02	-6.77E-02
trainer	2.54E-01	-4.44E-02
jogging	4.97E-01	-2.64E-02
fitbit	9.26E-01	-3.60E-03
ejercicios	9.78E-01	-1.07E-03

Figure 1 Feature Engineering: Correlation Analysis (a. Obesity; b. Exercise)

(1)	*Green:*	*|r|>0.4, p≤.05,*	*most informative,*	*for both humans and machine learning*
(2)	*Blue:*	*|r|≤0.4, p≤.05,*	*informative,*	*used in machine learning*
(3)	*Orange:*	*p>.05,*	*uninformative,*	*discarded*

Our correlation analysis, detailed in Figure 1, where parts (a) and (b) correspond to obesity and exercise respectively, followed a dual-metric approach involving p-values and Pearson correlation coefficients (r-values). Keywords with p-values above 0.05, represented in orange in the figure, were deemed statistically insignificant and thus excluded from our model. The remaining keywords, shown in blue and green, met our criteria for inclusion in the subsequent phases of model training.

Beyond conventional machine learning standards, our study incorporated an additional layer of stringency for feature selection based on human interpretability, setting an absolute r-value threshold of greater than 0.4 to identify features with moderate-to-high correlation. Keywords in green, meeting this heightened criterion, were classified as the most informative for both computational and human-centric analysis, particularly valuable in public health research and policy-making contexts. The blue-marked keywords, while less correlated, were also included in the machine learning models due to their informative nature. This dual-threshold strategy, combining statistical rigor with practical applicability, underscores the thoroughness and adaptability of our approach in feature selection, ensuring that the chosen features are not only statistically significant but also hold substantial real-world relevance.

Machine Learning Models: Training, Testing, and Evaluation

We then applied four distinct machine learning models — Lasso, Random Forests, Ridge, and Polynomial Regression — to predict obesity and exercise prevalence rates using the select Google search keywords identified in the feature engineering stage (referenced in Figure 1). These models were rigorously trained on a dataset spanning from 2005 to 2017, ensuring a comprehensive learning process based on historical trends. The testing phase was conducted on data from the year 2018, providing a robust evaluation of each model's predictive performance in a real-world scenario. The outcomes of this predictive modeling, including the Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) for both obesity and exercise rate predictions, are systematically delineated in Tables 1 and 2. This structured approach allowed us to compare the efficacy of each model in a quantifiable manner, offering insights into their respective strengths and limitations in the context of public health trend prediction.

Table 1: Comparative Performance of Machine Learning Models in Predicting Obesity Prevalence Rates

Obesity:		**Validation**			**Test**		
Model & Best Parameter		**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**
Lasso	alpha=0.001	3.17E-04	1.40E-02	7.66E-01	6.27E-04	2.07E-02	5.78E-01
Random Forest		3.89E-04	1.58E-02	7.13E-01	5.34E-04	1.92E-02	6.41E-01
Ridge		1.97E-04	1.11E-02	8.55E-01	3.91E-04	1.56E-02	7.37E-01
Polynomial	degree=1	4.04E-04	1.54E-02	8.60E-01	3.55E-04	1.51E-02	7.61E-01

Table 2: Comparative Performance of Machine Learning Models in Predicting Exercise Prevalence Rates

Exercise:		**Validation**			**Test**		
Model & Best Parameter		**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**
Lasso	alpha=0.001	4.94E-04	1.73E-02	7.22E-01	6.84E-04	2.16E-02	5.31E-01
Random Forest		6.31E-04	1.98E-02	6.44E-01	8.36E-04	2.42E-02	4.27E-01
Ridge		3.71E-04	1.53E-02	7.91E-01	3.46E-04	1.58E-02	7.62E-01
Polynomial	degree=3	6.49E-04	2.04E-02	5.31E-01	5.55E-04	1.81E-02	6.20E-01

We proceeded to model evaluation, adopting the Coefficient of Determination (R^2) as our primary performance metric. This choice was made to capture the model's effectiveness in explaining the variance of the dependent variables (obesity and exercise prevalence rates) from the independent variables (Google search keywords). Unlike Mean Squared Error (MSE) or Mean Absolute Error (MAE), R^2 provides a holistic measure of model performance, particularly beneficial when dealing with datasets that may include outliers or when MSE and MAE offer limited insights due to their inherent characteristics. MSE can overly penalize larger errors due to squaring, and MAE, though robust to outliers, may not sufficiently reflect the model's predictive capacity. By employing R^2, we aimed to achieve a more comprehensive understanding of how well our selected features, validated for their statistical significance, could explain the variability in the target variables. This approach underlines our commitment to ensuring a robust, interpretable, and scientifically sound evaluation of our model's ability to predict public health trends from digital trace data.

Regarding the Coefficient of Determination (R^2) in our study on obesity prediction models, the Lasso Regression model, with an alpha value of 0.001, achieved an R^2 of 7.66E-01 in the validation phase and 5.78E-01 in the test phase, while the Random Forest model exhibited an R^2 of 7.13E-01 during validation and 6.41E-01 in testing. Ridge Regression showed a notably higher performance with an R^2 of 8.55E-01 in the validation phase and 7.37E-01 in the testing phase. Finally, the Polynomial Regression model with a degree of 1 recorded an R^2 of 8.60E-01 in validation and 7.61E-01 in testing. These results highlight the varying degrees of success in different regression models in explaining the variance in obesity prevalence rates, with Polynomial Regression demonstrating the strongest explanatory power.

In predicting exercise prevalence rates, the Lasso Regression model with alpha set to 0.001 demonstrated an R^2 of 7.22E-01 in the validation phase and 5.31E-01 in the test phase. The Random Forest model showed an R^2 of 6.44E-01 during validation and 4.27E-01 in testing, indicating a relatively lower explanatory power. Ridge Regression exhibited a stronger performance with an R^2 of 7.91E-01 in validation and 7.62E-01 in testing, suggesting a high degree of model effectiveness. Lastly, the Polynomial Regression model with a degree of 3 achieved an R^2 of 5.31E-01 in the validation phase and 6.20E-01 in testing. These outcomes reveal the diverse capabilities of the models in capturing the variance in exercise prevalence rates, with Ridge Regression notably excelling in explaining the data variability.

In this study, we selected Random Forest as the baseline model to assess effective prediction strategies. The rationale for this choice and a detailed strengths and limitations analysis are provided in the Discussion section. Our results, illustrated in the accompanying chart, reveal that in predicting obesity, both Polynomial and Ridge Regression models outperform the baseline. The Lasso Regression model, however, does not show this improvement. The Polynomial Model with a degree of 1 exhibits the highest R-squared value, indicating superior accuracy. For exercise prediction, all models, particularly the Ridge Regression, demonstrate better performance than the baseline, with Ridge achieving the highest R-squared value. This highlights its effectiveness in capturing the variance within the exercise data set.

Discussion

In our study on US obesity and exercise trends from 2005 to 2018, the Random Forest model stands out as a benchmark for its effectiveness in diverse scenarios. Key to our objectives of rate prediction and trend identification, it skillfully handles regression and classification tasks. Its robust handling of data variability and noise over a decade enhances its suitability. The model’s capability to identify search trends closely correlated with obesity and exercise rates provides critical insights. Utilizing Random Forest as a baseline not only deepens our understanding of influential factors but also enables comparative analysis with other models, potentially revealing subtler insights or confirming our initial findings. This approach, leveraging Random Forest's versatility and robustness, keeps options open for exploring other models, ensuring a thorough analysis of trends in exercise participation and obesity.

Limitation

A significant limitation of our research is the absence of stratified data like gender and age in the Google search dataset. This lack of detail hampers our ability to analyze trends within specific demographic groups, limiting the depth of our conclusions about population segments' engagement with exercise and obesity topics. This constraint impacts our ability to fully represent the diversity of these trends across different demographic categories.

Strength

Despite these challenges, our study's strengths lie in its cost-effectiveness and timeliness. Analyzing publicly available search trend data using Random Forest models is an economical approach to understanding health behaviors on a large scale. This methodology, contrasting with traditional, more costly data collection methods, allows for more feasible and timely research, offering insights valuable for public health strategies and interventions.

References

Memon, Shahan Ali, Saquib Razak, and Ingmar Weber. “Lifestyle disease surveillance using population search behavior: Feasibility study.” Journal of Medical Internet Research 22.1 (2020): e13347. URL: https://www.jmir.org/2020/1/e13347/

Yousif, M.M., Kaddam, L.A. & Humeda, H.S. Correlation between physical activity, eating behavior and obesity among Sudanese medical students Sudan. BMC Nutr 5, 6 (2019). https://doi.org/10.1186/s40795-019-0271-1

Deurenberg, P., & Yap, M. (1999). The assessment of obesity: methods for measuring body fat and global prevalence of obesity. Bailliere's best practice & research. Clinical endocrinology & metabolism, 13(1), 1–11. https://doi.org/10.1053/beem.1999.0003

Acknowledgments

We want to sincerely thank our instructors, Prof. Bedoor K. AlShebli and Dr. Sara Saboor, for their invaluable guidance and insight, which was instrumental in the completion of this work.

Appendix

obesity_keywords = ['abdominal', 'calories', 'cholesterol', 'diabetes insipidus', 'diabetes insulin', 'diabetes mellitus', 'diabetes symptoms', 'diabetes', 'diabetic diet', 'diabetic ketoacidosis', 'diabetic', 'diet', 'dietary', 'dresses plus size', 'food delivery near me', 'gastric', 'glucose', 'glycogen', 'how to lose weight', 'hyperglycemia', 'hypertension', 'insulin', 'insulin syringes', 'ketoacidosis', 'malnutrition', 'meals', 'nutrition', 'obese', 'obesity', 'pizza delivery', 'polyphagia', 'prediabetes', 'quinoa gluten free', 'sclerosis', 'signs of diabetes', 'slim', 'sugar level', 'symptoms of congestive heart failure', 'symptoms of diabetes', 'symptoms of high blood sugar', 'type 2 diabetes', 'type 2', 'weighing', 'weight loss']

exercise_keywords = ['aerobic exercise', 'best workout', 'bike helmet', 'bike laws', 'bike locks', 'bike repair', 'bike sale', 'bodybuilding', 'e-bike', 'ejercicios', 'endocrine', 'exercise', 'exercises', 'fitbit', 'fitness gym', 'fitness', 'gym near me', 'gym', 'healthy', 'how to exercise', 'iPod', 'inactivity', 'insanity workout', 'ipod reset', 'jogging', 'my fitness pal', 'pre workout', 't25 schedule', 'trainer', 'wellness', 'workout', 'yoga']