DIGITAL
HEALTH

Leveraging online shopping behaviors as a proxy for personal lifestyle choices: New insights into chronic disease prevention literacy

Yongzhen Wang[1] (iD), Xiaozhong Liu[2], Katy Börner[3], Jun Lin[4], Yingnan Ju[3], Changlong Sun[4] and Luo Si[4]

Abstract

Objective: Ubiquitous internet access is reshaping the way we live, but it is accompanied by unprecedented challenges in preventing chronic diseases that are usually planted by long exposure to unhealthy lifestyles. This paper proposes leveraging online shopping behaviors as a proxy for personal lifestyle choices to improve chronic disease prevention literacy, targeted for times when e-commerce user experience has been assimilated into most people's everyday lives.

Methods: Longitudinal query logs and purchase records from 15 million online shoppers were accessed, constructing a broad spectrum of lifestyle features covering various product categories and buyer personas. Using the lifestyle-related information preceding online shoppers' first purchases of specific prescription drugs, we could determine associations between their past lifestyle choices and whether they suffered from a particular chronic disease.

Results: Novel lifestyle risk factors were discovered in two exemplars—depression and type 2 diabetes, most of which showed reasonable consistency with existing healthcare knowledge. Further, such empirical findings could be adopted to locate online shoppers at higher risk of these chronic diseases with decent accuracy [i.e. (area under the receiver operating characteristic curve) AUC=0.68 for depression and AUC=0.70 for type 2 diabetes], closely matching the performance of screening surveys benchmarked against medical diagnosis.

Conclusions: Mining online shopping behaviors can point medical experts to a series of lifestyle issues associated with chronic diseases that are less explored to date. Hopefully, unobtrusive chronic disease surveillance via e-commerce sites can grant consenting individuals a privilege to be connected more readily with the medical profession and sophistication.

Introduction

Chronic diseases, such as heart disease, cancer, and diabetes, are the leading causes of mortality and disability across the globe. For instance, in the United States, 6-in-10 adults live with at least one chronic disease, and 70% of deaths and nearly 75% of aggregate healthcare spending are due to chronic diseases.[1] Similarly, in China, the largest developing country with a population of 1.4 billion, chronic diseases account for an estimated 80% of deaths and 70% of disability-adjusted life years lost.[2] Although early detection of chronic

[1]Institute of Science of Science and S&T Management, Dalian University of Technology, Dalian, China
[2]Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA
[3]School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA
[4]Alibaba DAMO Academy, Hangzhou, China

Corresponding authors:
Yongzhen Wang, Institute of Science of Science and S&T Management, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian 116024, China.
Email: yongzhenwang@dlut.edu.cn

Xiaozhong Liu, Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA.
Email: xliu14@wpi.edu

diseases can trigger timely cures and prolonged survival, regular health checkups can be cost-prohibitive[3] and incentive-deficient.[4] To alleviate underdiagnosis and undertreatment, healthy lifestyles such as a balanced diet,[5] smoking cessation,[6] physical exercise,[7] and alcohol withdrawal[8] have long been advocated as preventive measures against chronic diseases. However, new avenues to granular, actionable data profiling for lifestyles toward the goal of chronic disease prevention are still needed, since booming digital platforms continuously revolutionize our everyday lives through the internet of everything.[9]

Recently, it was reported that predicting the early risk of chronic kidney disease in patients with diabetes via real-world data showed enhanced performance versus clinical ones.[10] In addition, similar findings were also drawn from a series of studies built upon social media sites, on which social media users' lifestyle changes could be documented unobtrusively. For example, Twitter corpora were used to develop a statistical classifier to provide a risk estimate of how depressed a user was;[11] Reddit posts were used to create models capable of automatically detecting anxiety disorders;[12] Instagram photos were used to establish useful psychological indicators to reveal predictive markers of depression;[13] Facebook languages were used to predict occurrences of depression-related events in users' medical records.[14] Unlike dedicated and strictly controlled medical research, these data-driven works did not hinge on any hypothesis about chronic diseases, stimulating new ideas to achieve lifestyle-oriented data profiling based on digital platforms.

This paper proposes leveraging online shopping behaviors as a proxy for personal lifestyle choices, aiming to explore unrevealed lifestyle risk factors and accordingly open an innovative way to chronic disease risk prediction. The reasons and motivations are as follows. Online shopping has had a profound impact on the ways people live their lives—the benefits of online shopping are becoming seemingly endless and have changed the culture and behaviors of shoppers everywhere.[15] Amid the flourishing era of e-commerce, the number of online shoppers worldwide has now reached 2.14 billion,[16] and the global online retail sales are expected to increase up to 5.4 trillion dollars by 2022.[17] Here comes the question: is there a connection between online shoppers and healthcare consumption? One Harris Interactive study demonstrates that well online healthcare consumers make up approximately 60% of the consumers searching for health information online—they search for preventive medicine and wellness information in the same way they look for news, stock quotes, and products; as for those newly diagnosed, they will search frenetically in the first few weeks following their diagnosis, and many of them will cast a wide net for medical help online.[18] So it is conceivable that the incorporation of online shopping behaviors into chronic diseases prevention will grow in practice and importance as more people go shopping online.

This paper selects China, the world's top retail market with 35.3% of sales taking place online,[19] as the testbed. Meanwhile, Alibaba, the bellwether of China's e-market with a 53.3% share and 758 million active users,[20] is chosen as the data source. Alibaba situates a great wealth of goods and services in the context of almost every aspect of Chinese living consumption, ranging from explicit features, such as food intake and entertainment spending, to implicit ones, such as body size and clothing preference. On this basis, this paper sheds the spotlights on two representative chronic diseases—depression and type 2 diabetes that affect and can be affected by personal lifestyle choices[21,22]—as two case studies. The reasons and motivations are as follows. In China, rapid social and economic changes quickened the pace of living and caused a general increase in psychological pressure and stress. Recently, the prevalence rate of depressive symptoms among Chinese adults was estimated to be 37.9%,[23] but the recognition rate of depression in Shanghai, a Tier-1 city of China, was only 21%, far below the world average.[24] Similarly, along with China's urbanization and rising living standards during the past decades, the prevalence rate of type 2 diabetes soared from less than 1% in the 1980s to dramatically more than 10%.[25] Unfortunately, more than 60% of Chinese adults with type 2 diabetes were unaware of their diagnosis.[26]

This paper addresses the following two research questions:

- Can online shopping behaviors reveal lifestyle risk factors associated with chronic diseases?
- Can online shopping behaviors be utilized for chronic disease risk prediction?

To the best of our knowledge, we have made the first attempt to translate online shopping behaviors to lifestyle-oriented data profiling in order to improve chronic disease prevention literacy. As one of the deliverables of this paper, we find that "if female online shoppers aged 15-to-24 used to (i) shop frequently, (ii) endure financial pressure, and (iii) spend more on healthcare products, alcoholic drinks, haircare services, and reading materials, but (iv) pay less for phone expenses and women's clothing, they could face a higher chance of developing depression." Comparably, such an empirical finding could be quite intractable to obtain for other digital platforms, even though some of them, such as Twitter, Reddit, Instagram, and Facebook, proved to identify online users with psychological abnormality via user-generated content, whose original intention was for social sharing.[11–14]

Materials and methods

Overview

This paper surveyed 15 million Alibaba users at random within a year span from 1 January to 31 December 2018; however, users were excluded if they had bought any prescription drug prior to 2018 to guarantee an unbiased analysis for the later onset of chronic diseases. The yearlong span was partitioned into an eight-month observation period (i.e. January through August) and a four-month performance period (i.e. September through December). Notably, only information generated within the observation period, including 6 billion query logs and 3.2 billion purchase records, were used as input for lifestyle-oriented data profiling. As for the two case studies, users who bought prescription drugs for depression and type 2 diabetes (shown in Figure 1 and listed in Supplemental Appendix A) during the performance period for the first time were defined as depressed and diabetic users, respectively. In contrast, those who did not make such purchases over the performance period were specified as control users. The sampling ratios of users with respect to depressed versus control and diabetic versus control were balanced at 1-to-19 and 1-to-9, respectively. These settings allowed us to compare different online shoppers' personal lifestyle choices across the same time window and simulate both prevalence rates of depression and type 2 diabetes in China (i.e. approximately 5% and 10%).[27,28]

Data collection

From 15 million Alibaba user accounts, we retrieved demographics (sex and age) and a total of 6 billion query logs, and 3.2 billion purchase records beginning on 1 January 2018 and ending on 31 December 2018. For the two case studies, 3071 depressed and 3936 diabetic users were identified. The user characteristics are reported in Table 1. Note that the control users (also the majority of the subjects of this research) reveal some basic characteristics of Chinese online shoppers. Overall, depressed and diabetic users are older (Student's t-test: for depression, $t' = 14.874$, $df = 4431.893$, $p < 0.001$; for type 2 diabetes, $t' = 34.899$, $df = 5687.344$, $p < 0.001$) and include a smaller proportion of women (Pearson's chi-squared test: for depression, $\chi^2 = 333.180$, $df = 1$, $N = 13071$, $p < 0.001$; for type 2 diabetes, $\chi^2 = 107.035$, $df = 1$, $N = 13936$, $p < 0.001$). Furthermore, they both tend to place more orders (Mann–Whitney U test: for depression, $U = 14,661,834$, $z = -3.790$, $p < 0.001$; for type 2 diabetes, $U = 17,241,491$, $z = -11.406$, $p < 0.001$), while diabetic users tend to search for products less frequently (Mann–Whitney U test: for depression, $U = 15,013,757.5$, $z = -0.283$, $p = 0.777$ for type 2 diabetes, $U = 18,322,854.5$, $z = -4.472$, p

< 0.001). In order to model the two case studies in a real-world scenario, for each depressed user, we randomly selected another 19 control users of the same demographics, yielding a sample of $3071 + 19 \times 3071 = 61,420$ users (i.e. 1-to-19 for depressed vs. control) for investigating depression.[27] Analogously, a sample of $3936 + 9 \times 3936 = 39,360$ users (i.e. 1-to-9 for diabetic vs. control) was chosen for investigating type 2 diabetes.[28] Each sample was then divided into several parts for subgroup analysis, keeping depressed/diabetic users and their control counterparts within the same subsample, given the fact that online shopping behaviors differed significantly by sex and age (explained in Supplemental Appendix B).

Feature engineering and selection

Two types of lifestyle features were engineered. One type was to unfold online shoppers' daily purchases explicitly (e.g. food intake and entertainment spending), while another type was to characterize their living consumption implicitly (e.g. body size and clothing preference). More concretely, 135 explicit lifestyle features were generated by pooling 3.2 billion purchase records into a list of sales statistics according to Alibaba product categorization. Meanwhile, 115 implicit lifestyle features were constructed by creating a variety of buyer personas from 6 billion query logs and 3.2 billion purchase records using Alibaba off-the-shelf data mining technologies. Notably, buyer personas for Alibaba users under 18 years old ($\approx 0.5\%$ of the subjects of this research) were not generated due to specific data policy. All these features were discretized to group data into different bins, and we prudently eliminated the collinearity among them. As for the two case studies, only the lifestyle features (listed in Supplemental Appendix C) showing a correlation with depression/type 2 diabetes in the chi-squared test of independence (quantified by a humble significance threshold $p < 0.1$) were retained for further analysis.

Regression analysis

Consider a regression model with multiple explanatory variables $x_1, x_2, \ldots, x_m$ and one binary explained variable y, aiming to estimate the probability of online shoppers suffering from a particular chronic disease $\pi = \Pr(y = 1 | x_1, x_2, \ldots, x_m)$. For clarity, $x_1, x_2, \ldots, x_m$ represents an array of m lifestyle features of an online shopper, and $y = 1$ indicates that he/she will be diagnosed with the focal chronic disease in the future (otherwise $y = 0$). Without loss of generality, a linear relationship is assumed between $x_1, x_2, \ldots, x_m$ and the log odds of $y = 1$, and multiple logistic regression[29] can be established as follows:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Here, the odds ratio (OR) is used to interpret the standardized regression coefficients $\beta_1, \beta_2, \ldots, \beta_m$. For example,
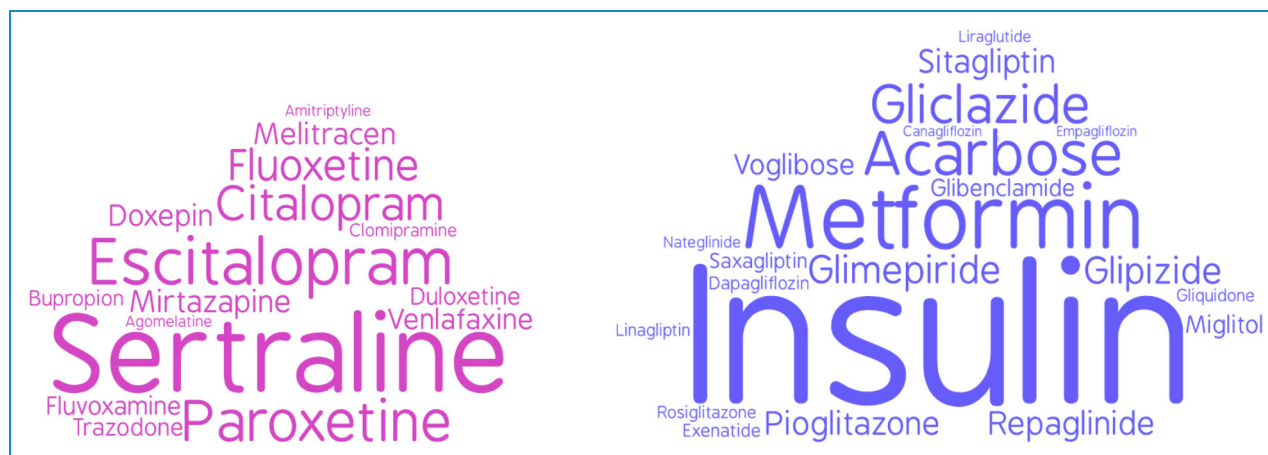
Figure 1. Prescription drugs found in Alibaba users' purchase records that are relevant to depression (*left*) and type 2 diabetes (*right*). Here, the generic name of each prescription drug is listed for simplicity. Note that larger fonts indicate higher relative purchase frequencies.

Table 1. Online shopper characteristics. Here, 10,000 control (Alibaba) users were randomly sampled for ease of comparison. Differences in age were tested by the Student's *t*-test with equal variances not assumed, percent female by the Pearson's chi-squared test, query count, and purchase count by the Mann−Whitney *U* test, in comparison with the control users.

Descriptive	Depressed	Diabetic	Control
#Subject	3071	3936	10000
Age (M ± SD)	34.4 ± 10.5*	38.5 ± 11.8*	31.3 ± 8.7
Female (%)	37.9*	47.0*	56.7
Monthly #Query (M ± SD)	48.9 ± 58.5	44.5 ± 50.4*	46.6 ± 50.3
Monthly #Purchase (M ± SD)	13.8 ± 11.8*	15.9 ± 13.5*	13.1 ± 12.2

M: mean; SD: standard deviation.
*Significant difference ($\alpha = 0.05$).

the *OR* for a one unit increase in x_1 is $\exp(\beta_1)$—there is a $[\exp(\beta_1) - 1] \times 100\%$ increase or decrease in the odds of $y = 1$ when x_1 increases by one unit. If $\exp(\beta_1) > 1$, then x_1 is positively (or negatively if $\exp(\beta_1) < 1$) associated with the odds of $y = 1$.

Statistical power analysis

A priori power analysis was performed for subsample size estimation. The effect size in this paper was expected to be $OR = 1.49$ (or inverted 0.67), considered to be small using Cohen's criteria.[30] With $\alpha = 0.05$, power $= 0.8$, probability of null hypothesis being true $= 0.5$, proportion of variance for other covariates $= 0.8$, the projected subsample size needed with this effect size was approximately 1068 for logistic regression. Therefore, we excluded the depression subsamples aged

above 55 and type 2 diabetes subsamples aged above 65 due to insufficient statistical power, as shown in Table 2. Consequently, only 59,140 ($\approx$96.3%) and 38,420 ($\approx$97.6%) Alibaba users were retained for the two case studies of depression and type 2 diabetes, respectively.

Remarkably, for type 2 diabetes, the subsamples aged above 55 make up 10.42% of the total, quite close to the share of online shoppers over 50 years old in China (i.e. 10.7%).[31] But for depression, the subsamples aged above 55 account for only 3.71% of the total. In addition, the age distribution of online shoppers varies significantly from country to country. Take the United States as an example, 29% of online shoppers are 55 and older,[32] almost three times the number in China.

Table 2. Subsample statistics. Note that the subsamples with a strikethrough are removed due to insufficient statistical power.

Chronic disease	Sex	Age						Total
		15–24	25–34	35–44	45–54	55–64	65–74	
Depression	Female	5660	6500	6200	3900	~~920~~	~~80~~	59,140
	Male	7400	13,840	10,300	5340	~~920~~	~~360~~	
Type 2 diabetes	Female	2960	6400	4940	2870	1160	~~160~~	38,420
	Male	1600	5520	5930	5040	2000	~~780~~	

Multiple comparisons

In this paper, a lifestyle feature was considered a lifestyle risk factor if its estimated regression coefficient was significantly different from zero, that is, if it was significantly associated with chronic disease onset. The Benjamini–Hochberg (BH) adjusted *p*-value[33] was then applied to control false-positive errors among the discovered lifestyle risk factors at a level below 5%.

Prediction procedure

We employed a support vector machine[34] with a misclassification cost parameter (illustrated in Supplemental Appendix D) to develop predictive classifiers to provide risk estimates for chronic diseases. All selected lifestyle features were used as predictors. In practice, we deliberately chose the area under the receiver operating characteristic curve (AUC), a widely used indicator suitable for describing the classification accuracy over imbalanced classes, as our evaluation metric.[35] Moreover, 10-fold cross-validation was adopted to avoid over-fitting as follows. Each subsample was partitioned into 10 stratified folds—one predictive classifier was trained using nine folds and was evaluated using the remaining held-out fold. This process was repeated 10 times, each time with a different held-out fold, and then the results were averaged, yielding cross-validated out-of-sample AUC for performance assessment.

Exploration of lifestyle risk factors

To reveal associations between online shoppers' past lifestyle choices and whether they suffered from a particular chronic disease or not, we first mapped their historical query logs and purchase records into a wide array of interpretable and fine-grained lifestyle features, then chose multiple logistic regression[29] as the analytical process allowing for an explanation of this data-driven exploration (detailed in "Regression analysis" subsection under "Materials and methods" section). Here, the lifestyle features significantly associated with chronic disease onset can be regarded as lifestyle risk factors. Figure 2 shows all lifestyle risk factors derived from the two case studies of depression and type 2 diabetes when demographics are controlled for. More information about the regression output can be found in Supplemental Appendix E.

Case 1: depression

For depression, the distribution of discovered lifestyle risk factors presents a conspicuous downtrend with the aging of online shoppers, roughly aligning with the distribution of ages at the onset of the first major depressive episode.[36] Understandably, the shopping category most closely associated with depression falls on healthcare products such as over-the-counter drugs and contraceptives (e.g. *home healthcare supplies* $@_{15-24}^{\text{female}}$, $OR = 1.570$, $\beta = 0.451 \pm 0.079$, $p^{\text{BH}} < 0.001$).

Overall, depressed online shoppers from 15-to-54 years old manifest a stronger user viscosity to e-commerce sites than otherwise healthy counterparts (e.g. *purchase frequency* $@_{15-24}^{\text{male}}$, $OR = 1.173$, $\beta = 0.159 \pm 0.030$, $p^{\text{BH}} < 0.001$; *purchase amount* $@_{25-34}^{\text{female}}$, $OR = 1.195$, $\beta = 0.178 \pm 0.051$, $p^{\text{BH}} = 0.004$). Reasonably, e-commerce platforms can provide a superior shopping venue for individuals with depression potentials who are less proactive to participate in face-to-face social interactions.[37] Further, 15-to-24 year-old depressed online shoppers are inclined to show special interest in reading materials (e.g. *books, magazines & newspapers* $@_{15-24}^{\text{female}}$, $OR = 1.171$, $\beta = 0.158 \pm 0.055$, $p^{\text{BH}} = 0.021$). Another impressive observation lies in that female depressed online shoppers aged 25-to-34 tend to buy clothing for the opposite sex and children less frequently (e.g. *men's clothing* $@_{25-34}^{\text{female}}$, $OR = 0.681$, $\beta = -0.384 \pm 0.141$, $p^{\text{BH}} = 0.038$; *children's clothing* $@_{25-34}^{\text{female}}$, $OR = 0.856$, $\beta = -0.155 \pm 0.060$, $p^{\text{BH}} = 0.043$), while male depressed online shoppers aged 25-to-44 are less prone to pay their attention to kids' clothes and home improvements (e.g. *children's clothing* $@_{25-34}^{\text{male}}$, $OR = 0.704$, $\beta = -0.351 \pm 0.090$, $p^{\text{BH}} < 0.001$; *home decoration preference* $@_{25-34}^{\text{male}}$,
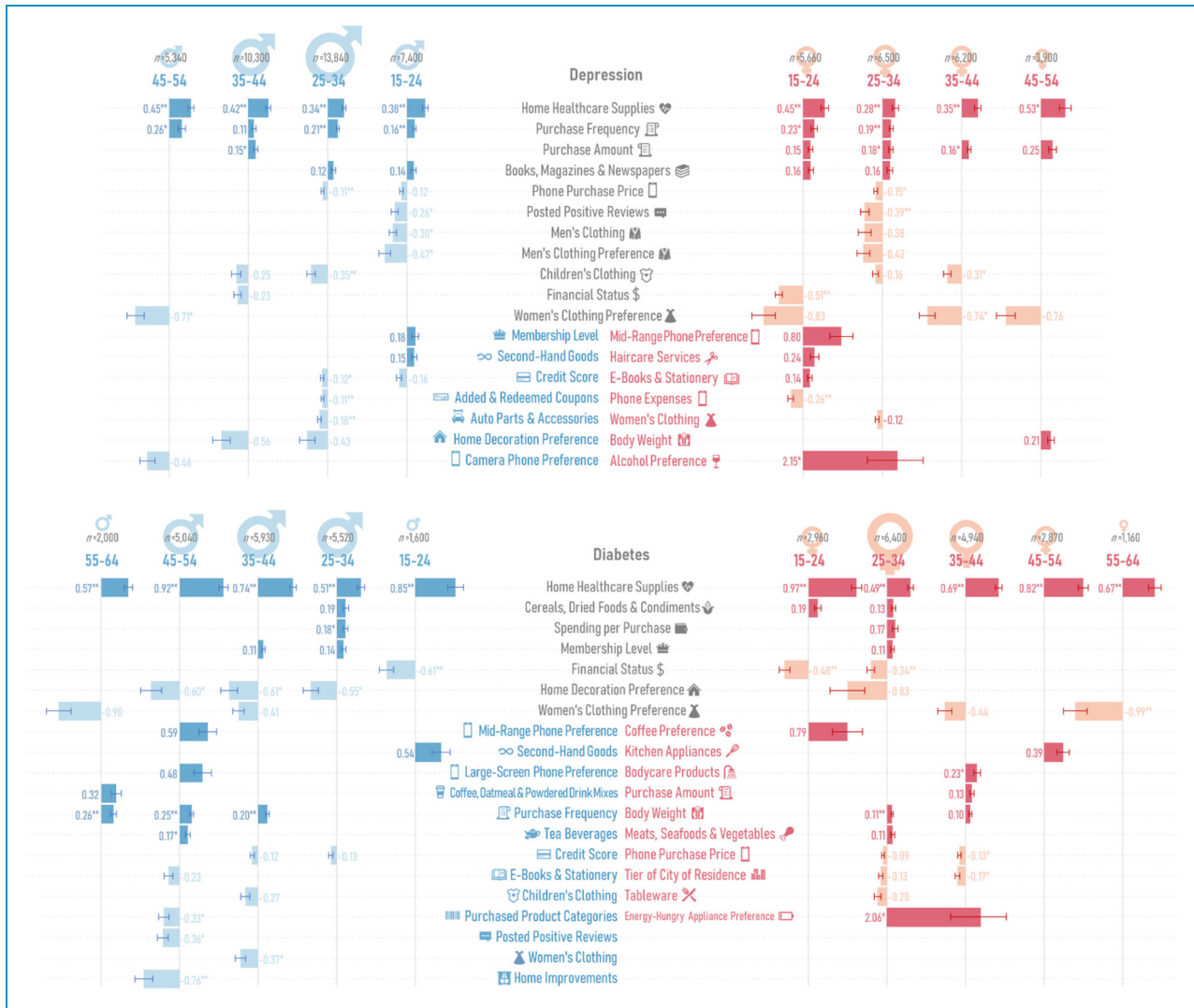
Figure 2. Discovered lifestyle risk factors associated with depression (*top*, $N = 59{,}140$) and type 2 diabetes (*bottom*, $N = 38{,}420$) when demographics are controlled for, reported as standardized regression coefficients (bars) and standard errors (two-end lines). Note that each column represents a subgroup analysis using multiple logistic regression, with the related sex, age, and size n of each subsample marked on top. Multiple significance tests are conducted by the Benjamini–Hochberg (BH) procedure. $*p^{BH} < 0.01$, $**p^{BH} < 0.001$, otherwise $p^{BH} < 0.05$.

$OR = 0.649$, $\beta = -0.432 \pm 0.163$, $p^{BH} = 0.048$). To a significant extent, these depressed online shoppers are more likely to be single, moderately supporting the traditional perspective that marriage or cohabiting can promote psychological well-being.[38]

There are some other findings consistent with the existing literature on depression. For example, a glaring positive association stands between depression and alcohol consumption among female online shoppers of 15- to 24-years old (*alcohol preference* $@_{15-24}^{\mathrm{female}}$, $OR = 8.612$, $\beta = 2.153 \pm 0.598$, $p^{BH} = 0.005$), reiterating the causal inference that increased alcohol involvement raises the incidence rate of depression for adolescents and youth.[39]

Moreover, female depressed online shoppers aged 15-to-24 are more likely to suffer from hair problems (*haircare services* $@_{15-24}^{\mathrm{female}}$, $OR = 1.268$, $\beta = 0.237 \pm 0.089$, $p^{BH} = 0.032$), confirming the popular perception that hair damage and mental disorders often occur in combination.[40] As for female depressed online shoppers at the age of 45-to-54, they can be exposed to greater risk of being overweight (*body weight* $@_{45-54}^{\mathrm{female}}$, $OR = 1.239$, $\beta = 0.214 \pm 0.071$, $p^{BH} = 0.026$), matching the bi-directional relationship between depression and obesity in middle-aged women.[41] In addition, several discovered lifestyle risk factors relevant to personal finances (e.g. *financial status* $@_{15-24}^{\mathrm{female}}$, $OR = 0.602$, $\beta = -0.507 \pm 0.074$, $p^{BH} < 0.001$; *credit score*

$@_{15-24}^{\text{male}}$, $OR = 0.849$, $\beta = -0.164 \pm 0.053$, $p^{\text{BH}} = 0.015$) conform to the subjective intuition that lacking money usually leads to magnified feelings of anxiety and depression in many people.[42]

Case 2: type 2 diabetes

For type 2 diabetes, the number of discovered lifestyle risk factors for female online shoppers and their male counterparts peaks at the age of 25-to-34 and 45-to-54, respectively. This discrepancy partly echoes the sex difference that women have a higher prevalence of type 2 diabetes in youth while men see a higher prevalence in midlife.[43]

Just like in the case of depression, the most salient shopping category associated with type 2 diabetes points to healthcare products (e.g. *home healthcare supplies* $@_{15-24}^{\text{female}}$, $OR = 2.639$, $\beta = 0.970 \pm 0.105$, $p^{\text{BH}} < 0.001$). Also, a majority of diabetic online shoppers display high consumer inertia with respect to e-commerce sites (e.g. *membership level* $@_{25-34}^{\text{female}}$, $OR = 1.118$, $\beta = 0.112 \pm 0.042$, $p^{\text{BH}} = 0.043$; *purchase frequency* $@_{35-44}^{\text{male}}$, $OR = 1.219$, $\beta = 0.198 \pm 0.037$, $p^{\text{BH}} < 0.001$). On the whole, however, diabetic online shoppers favor much more diet-related purchasing activities (e.g. *cereals, dried foods & condiments* $@_{15-24}^{\text{female}}$, $OR = 1.205$, $\beta = 0.187 \pm 0.065$, $p^{\text{BH}} = 0.047$; *coffee, oatmeal & powdered drink mixes* $@_{55-64}^{\text{male}}$, $OR = 1.377$, $\beta = 0.320 \pm 0.106$, $p^{\text{BH}} = 0.012$). Clearly, food intake patterns play a crucial role during the progression toward type 2 diabetes.[44] In addition, female diabetic online shoppers from 25-to-34 and 45-to-54 years old exhibit a unique preference for household appliances (*energy-hungry appliance preference* $@_{25-34}^{\text{female}}$, $OR = 7.862$, $\beta = 2.062 \pm 0.663$, $p^{\text{BH}} = 0.003$; *kitchen appliances* $@_{45-54}^{\text{female}}$, $OR = 1.484$, $\beta = 0.395 \pm 0.127$, $p^{\text{BH}} = 0.031$).

There exist some other findings congruent with previous studies on type 2 diabetes. For example, 15-to-34 year-old diabetic online shoppers tend to endure financial pressure (e.g. *financial status* $@_{15-24}^{\text{female}}$, $OR = 0.617$, $\beta = -0.483 \pm 0.106$, $p^{\text{BH}} < 0.001$; *credit score* $@_{25-34}^{\text{male}}$, $OR = 0.881$, $\beta = -0.127 \pm 0.045$, $p^{\text{BH}} = 0.039$), corroborating the social inequality that poverty enlarges the likelihood of developing type 2 diabetes.[45] Another plausible conclusion can be that a larger proportion of female diabetic online shoppers at the ages of 25-to-44 may settle in big cities (e.g. *tier of city of residence* $@_{35-44}^{\text{female}}$, $OR = 0.841$, $\beta = -0.173 \pm 0.051$, $p^{\text{BH}} = 0.009$), aligning with the sharp increase in China's type 2 diabetes prevalence rate attached to its rapid diffusion of urbanization.[46] Moreover, female diabetic online shoppers aged 25-to-44 can be peculiarly susceptible to obesity (e.g. *body weight* $@_{25-34}^{\text{female}}$, $OR = 1.111$, $\beta = 0.105 \pm 0.023$, $p^{\text{BH}} < 0.001$), corresponding to the fact that excessive weight gain is a typical expression emanating from the progression of type 2 diabetes.[47] As for male diabetic online shoppers aged 35-to-44, fewer purchasing activities are observed when it comes to women's and kids' wear and home decoration (e.g. *women's clothing* $@_{35-44}^{\text{male}}$, $OR = 0.688$, $\beta = -0.374 \pm 0.107$, $p^{\text{BH}} = 0.004$; *children's clothing* $@_{35-44}^{\text{male}}$, $OR = 0.764$, $\beta = -0.269 \pm 0.098$, $p^{\text{BH}} = 0.029$; *home decoration preference* $@_{35-44}^{\text{male}}$, $OR = 0.542$, $\beta = -0.612 \pm 0.176$, $p^{\text{BH}} = 0.002$), suggesting that they are more likely to remain single. This outcome adds empirical evidence to the recent research revealing that marriage can, in a way, protect middle-aged and older men from adult-onset diabetes.[48]

Executive summary

This section presents a data-driven workflow to explore lifestyle risk factors underlying online shopping behaviors, with the purpose of improving chronic disease prevention literacy to cope with the ongoing digitalization of daily life. The discovered lifestyle risk factors involve a variety of product categories and buyer personas, most of which demonstrate reasonable consistency with existing healthcare knowledge. Also, these empirical findings can, to a certain degree, allow medical experts to capture the dynamics of chronic disease severity across time with a richness that is unavailable to conventional health checkups delivered at discrete points of time. More importantly, mining online shopping behaviors can point medical experts to a series of lifestyle issues associated with chronic diseases that are less explored to date.

Chronic disease risk prediction

To identify online shoppers at higher risk of chronic diseases, we built a support vector machine (SVM)[34] as the predictive classifier based on their past lifestyle choices that were reflected in their historical query logs and purchase records, along with the use of 10-fold cross-validation to avoid over-fitting (detailed in the "Prediction procedure" subsection under the "Materials and methods" section). This model employed the interpretable and fine-grained lifestyle features as predictors, and was evaluated by comparing the estimated probability of online shoppers suffering from a particular chronic disease against the actual presence or absence of related prescription drugs in their purchase records. By varying the threshold of predicted probabilities for classification, a receiver operating characteristic (ROC) curve was uniquely determined. The area under the ROC curve (AUC) was calculated as a proxy for the accuracy of the early risk prediction of chronic diseases.[35]

Performance analysis

Figure 3 illustrates the prediction performance for the two case studies of depression and type 2 diabetes when demographics are controlled for. Notably, all predictors here are divided into two disjoint subsets to validate the discovered
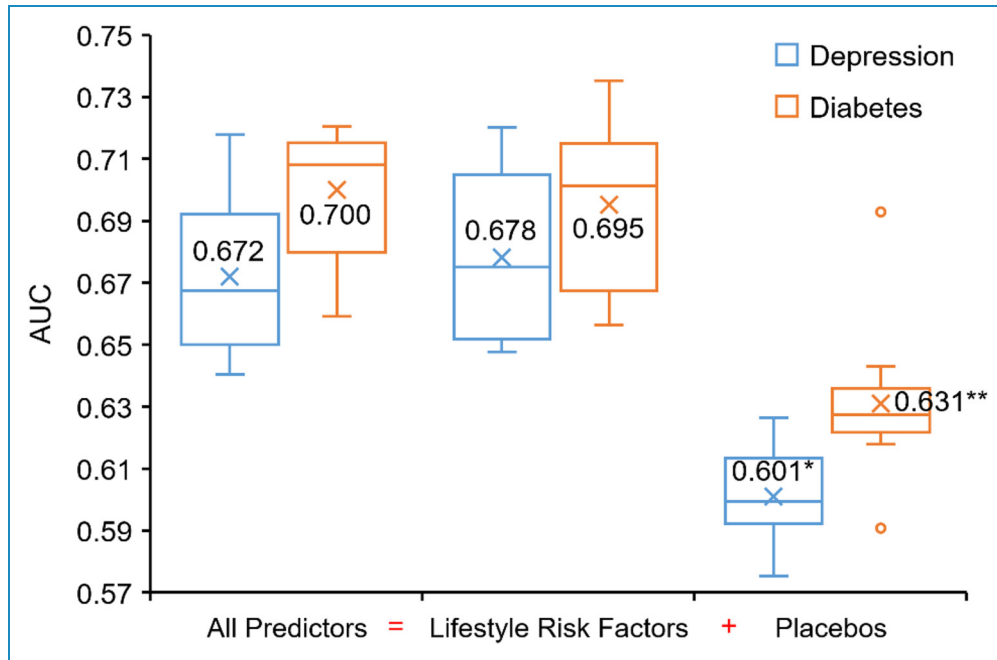
Figure 3. Online shopping behaviors-based prediction performance (via SVM) in early risk of depression ($N = 8$) and type 2 diabetes ($N = 10$) when demographics are controlled for. Statistical analysis is conducted by the Wilcoxon signed-rank test, in comparison with the lifestyle risk factors. $*p < 0.05$, $**p < 0.01$.
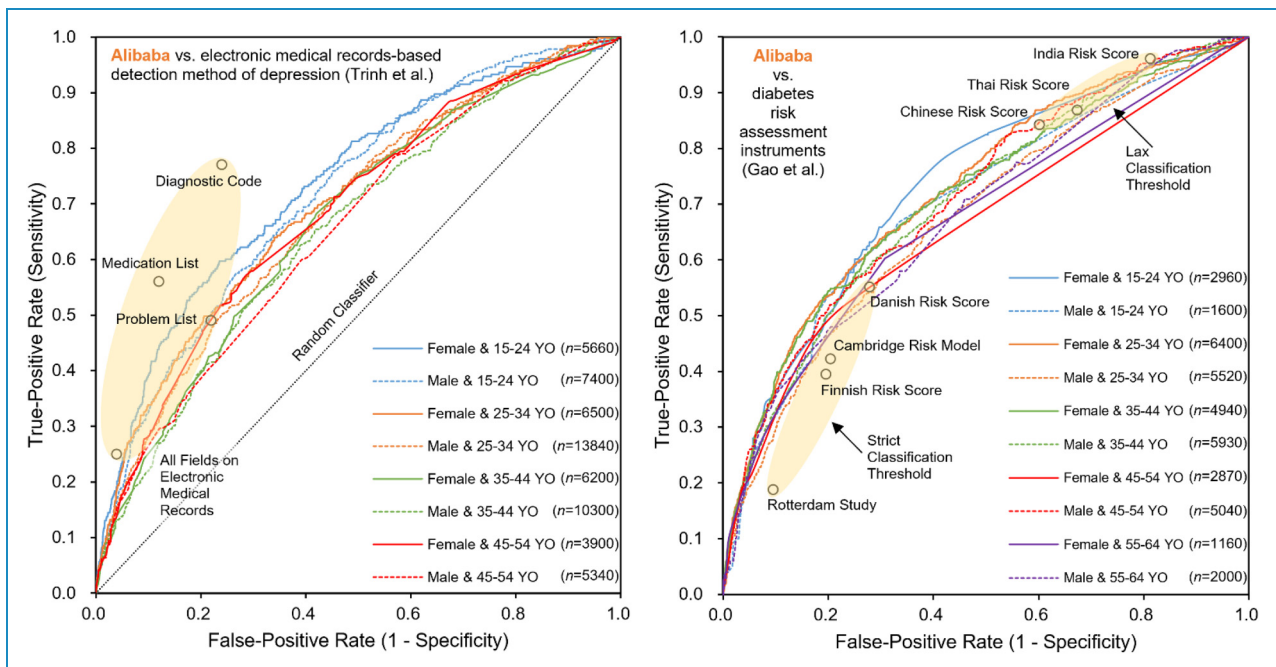AUC: area under the receiver operating characteristic curve; SVM: support vector machine.

Figure 4. In-sample ROC curves of online shopping behaviors-based predictive classifiers (via SVM) for depression (*left*, $N = 59,140$) and type 2 diabetes (*right*, $N = 38,420$) when demographics are controlled for. The points as combinations of true- and false-positive rates (shaded areas) are reported by the previous screening surveys—an electronic medical records-based detection method of depression ($N = 427$) and several diabetes risk assessment instruments ($N = 4336$).
ROC: receiver operating characteristic; YO: years old; SVM: support vector machine.

lifestyle risk factors empirically. On average, these lifestyle risk factors can result in cross-validated out-of-sample AUC of 0.678 and 0.695 when applied to depression and type 2 diabetes risk prediction, respectively, falling just short of the customary threshold for good discrimination (i.e. 0.7). Meanwhile, their predictive power significantly outperforms that of the placebos (Wilcoxon signed-rank test: for depression, $z = -2.521$, $p = 0.012$; for type 2 diabetes, $z = -2.803$, $p = 0.005$), and combining both subsets cannot improve final predictive accuracy by a substantial margin (Wilcoxon signed-rank test: for depression, $z = -1.260$, $p = 0.208$; for type 2 diabetes, $z = -1.174$, $p = 0.241$). To sum up, the discovered lifestyle risk factors can capture most of the depression-related or type 2 diabetes-related variance rooted in online shopping behaviors. Consequently, the placebos will be eliminated from all predictors to reduce model complexity. For more results about the prediction procedure, refer to Supplemental Appendix F.

Comparison with well-established screening surveys

Figure 4 compares our proposed online shopping behavior-based predictive classifiers against existing screening surveys for depression and type 2 diabetes, including an electronic medical records (EMRs)-based detection method of depression[49] and several diabetes risk assessment instruments[50] (detailed in Supplemental Appendix G). These baselines all select medical diagnosis as the gold standard for benchmarking. For depression, our proposed predictive classifiers perform nearly as well as those based on "diagnostic code," "problem list," and "medication list" jointly (i.e. three fields on EMRs) when a low false-positive rate is required. Even when it comes to a more relaxed restriction on false-positive errors, they still match closely with the baseline resorting solely to "problem list." Notably, the EMRs collected by Trinh et al.[49] originate from primary care patients, whereas the online shopping behaviors of this study come from a general population. As for type 2 diabetes, our proposed predictive classifiers display an obvious advantage over four Western-oriented risk prediction models (for American, Danish, Dutch, and Finnish) regarding relatively strict classification thresholds (i.e. a high cut-off probability for classifying a subject as positive) and, meanwhile, yield performance similar to three Asian-oriented ones (for Chinese, Indian, and Thai) regarding fairly lax classification criteria. Moreover, each of the seven baselines tested by Gao et al.[50] contains some strong predictors such as the family history of diabetes[51] and known condition of hypertension.[52] On the contrary, no hypothetical guidance has been applied to the feature engineering and selection of this research. Therefore, it is convincing to conclude that our proposed online shopping behaviors-based predictive classifiers can provide risk estimates for depression and type 2 diabetes as accurately as screening surveys benchmarked against medical diagnosis.

Executive summary

This section elaborates on how to pre-screen online shoppers for the prevention of chronic diseases by leveraging their "lifestyle profiles" documented in their historical query logs and purchase records. The growing population of online shoppers can thus be empowered to access low-cost early risk prediction of chronic diseases and accordingly jump-start personalized health intervention at the proper time—for example, e-commerce platforms should lessen or stop advertising sugary cereals and sweetened drinks to online shoppers who have already been informed of a high chance of heart disease. Hopefully, unobtrusive chronic disease surveillance via e-commerce sites is expected to be available for consenting individuals to be connected more readily with essential medical resources, cooperating with professional treatments and nursing to attain more guaranteed wellness.

Conclusions and outlook

Significance statement

Digitalization of daily life calls for new insights into chronic disease prevention literacy. This paper shows that online shopping behaviors can be leveraged as a proxy for personal lifestyle choices to discover novel lifestyle risk factors and provide accurate risk estimates for chronic diseases. It details relevant data mining workflows and results for two exemplars—depression and type 2 diabetes, arguing for translating online shoppers' historical query logs and purchase records to lifestyle-oriented data profiling which is akin to health checkups. Hopefully, unobtrusive chronic disease surveillance via e-commerce sites may soon meet consenting individuals in the digital space they already inhabit.

Contributions and implications

The core contribution of this paper lies in the fact that it links online shopping behaviors, a sound proxy for personal lifestyle choices, with chronic disease prevention, a public health challenge of paramount importance. In the two case studies, the discovered lifestyle risk factors are like fresh blood to research communities, including a great wealth of product categories and buyer personas, most of which exhibit reasonable consistency with the determinants and consequences of depression and type 2 diabetes. Further, these lifestyle risk factors manifest promising predictive power to serve as a scalable front-line alarm offering initial detection of depression and type 2 diabetes to give data-driven decision supports to medical practices. Our

experimental results suggest that online shopping behaviors documented in longitudinal query logs and purchase records should be integrated into current modalities for lifestyle-oriented data profiling, especially in today's digital era when social-, mobile-, and local-friendly e-commerce marketing are penetrating people's everyday lives far more easily and profoundly.[53] Along with the continuous improvement of data mining technologies, computational behavioral science in assistance with data profiling for lifestyles may become a dominant methodological paradigm for chronic disease prevention.

This paper also puts forward a critical concern about online shoppers' privacy breaches and informed consent for sharing personal data.[54] Realistically, some people, especially those who have to pay for their health insurance, can be unwilling to share their data for fear of disclosing health issues to any third party. Therefore, health administrators and policymakers should take special care to establish an ethical and supervised information exchange between healthcare networks and e-commerce systems, where online shopping behaviors are supposed to be considered protected health information subject to strict accessibility guidelines. Meanwhile, online shoppers should fully understand the secondary use of their historical query logs and purchase records—what will happen to their data, how their data will be used, and with whom their data will be shared—and maintain autonomous rights in their own healthcare decision-making. As health informatics progresses, unobtrusive chronic disease surveillance via e-commerce sites can be extended further to combine with other digital platforms that mirror personal lifestyle choices, such as social media sites like Twitter and Facebook, to improve health screening and help consenting individuals auto-complete self-report inventories to be assessed by medical experts in case of need.

Last but not least, this paper initiates a new perspective for the social responsibility fulfillment of online retailers in terms of public health—how to run a trade-off between user experience and user wellness. Obviously, e-commerce platforms can employ advanced data mining technologies to customize product recommendations to easily satisfy what online shoppers desire. However, in the long term, some product recommendations driven by customer preferences may irreparably impair online shoppers' physical abilities, especially for the vulnerable with chronic disease potentials. For example, it will cause serious health issues and even premature deaths if online retailers continue advertising sugary cereals and sweetened drinks to "sweet-toothed" online shoppers who have already been informed of a high chance of heart disease. Nevertheless, it is not in the interest of online retailers to stop such harmful advertisement—we have witnessed this in many cases with tobacco, formula milk for babies, etc. Necessary government interventions, such as requiring health warnings to be delivered before e-marketing, should be taken to achieve a win–win

situation—online shoppers improve health security, while e-commerce companies increase sustainable profitability.

Limitation and future work

The findings of this paper should be interpreted with caution due to the following limitation. In this study, whether online shoppers were labeled as "positive" (i.e. diagnosed with a particular chronic disease) or not entirely depended on the presence or absence of related prescription drugs in their purchase records. This practice could be problematic when so-called "positive" online shoppers searched for and bought prescription drugs just for their family members. Similarly, online shoppers who had never made such queries and purchases did not necessarily mean that they were by no means "positive" because they still could receive prescription drugs from offline pharmacies. Note that there are diabetes cases that are controlled through diet and physical activity, and there are off-the-counter medicines and non-pharmacologic treatments for depression. Furthermore, despite the increasing cases of type 2 diabetes in China (and elsewhere), the prevalence among teenagers and young adults is still relatively small. There is undeniable misinformation in and about the regression analysis of type 2 diabetes—some diabetes-related medications purchased by "positive" online shoppers at a young age can be used to treat other illnesses. For example, the association with haircare products, wigs, etc. in younger females may be due to the use of metformin as a treatment for polycystic ovary syndrome (PCOS)—hair loss is a symptom of PCOS. One more concern, from the statistical perspective, lies in the fact that some discussion points may go a little beyond what the data tell us, such as (i) supposing that all purchases are made for online shoppers themselves and (ii) making assumptions about online shoppers' circumstances and link this to a very heteronormative view of marriage and children. However, most of the findings based on such inferences can echo prior studies to a certain degree, and we have made an ambitious attempt to explore the possibility of utilizing online shopping history to explain and understand social phenomena.

This paper does not intend to compete with existing healthcare boards, but fills glaring gaps, from the individual level—empowering digital visitors and residents through near-real-time risk estimates for chronic diseases and situational awareness of lifestyle risk factors affecting their behavioral inertia, to the management level—contextualizing chronic disease prevention relative to evolving landscapes of the ongoing digitalization of daily life. In future work, an appealing direction would be to adopt medical diagnoses, such as the International Classification of Diseases codes from consenting individuals,[55] as the ground truth for health status assessment for online shoppers, in order to improve the representativeness of "positive" instances of chronic diseases. However, the ethical

issues associated with informed consent for data sharing should be reiterated—potential participants need to be informed of what would happen to their data, how their data would be used, and with whom their data would be shared. Another promising line of research would be to incorporate advanced technologies such as deep learning into current data mining workflows to reach fancier knowledge discovery. In addition, we would also like to generalize the proposed approach to investigate as many chronic diseases as possible to expand the horizon of digital health literacy.

Data sharing

Since Alibaba takes concerns of data privacy seriously, we would like to emphasize that none of the query logs and purchase records in this research's database permits specific identification with a particular individual and that the database retains no information about the identity, internet protocol address, or specific physical location of any user. At Alibaba, the query logs and purchase records are considered users' privacy and cannot be shared. However, for the discovered lifestyle risk factors, we can share their means and standard deviations (for ordinal coding) or categorical distributions (for nominal coding) with respect to depressed, diabetic, and control users. The data details are at https://doi.org/10.5281/zenodo.4722474.

ORCID iD: Yongzhen Wang https://orcid.org/0000-0001-7306-1291

References

1. Raghupathi W and Raghupathi V. An empirical study of chronic diseases in the United States: A visual analytics approach to public health. *Int J Environ Res Public Health* 2018; 15: 431.
2. Wang L, Kong L, Wu F, et al. Preventing chronic diseases in China. *Lancet* 2005; 366: 1821–1824.
3. Suhrcke M, Nugent RA, Stuckler D, et al. Cost-effectiveness of interventions to prevent chronic diseases. In: Neuschwander H (ed) *Chronic disease: An economic perspective*. London, UK: Oxford Health Alliance, 2006, pp.40–47.
4. Chien SY, Chuang MC, Chen I, et al. Primary drivers of willingness to continue to participate in community-based health screening for chronic diseases. *Int J Environ Res Public Health* 2019; 16: 1645.
5. National Research Council (US) Committee on Diet and Health. Impact of dietary patterns on chronic diseases. In: Morris R (ed) *Diet and health: Implications for reducing chronic disease risk*. Washington, DC: National Academies Press (US), 1989, pp.527–648.
6. Asaria P, Chisholm D, Mathers C, et al. Chronic disease prevention: Health effects and financial costs of strategies to reduce salt intake and control tobacco use. *Lancet* 2007; 370: 2044–2053.
7. Booth FW, Roberts CK and Laye MJ. Lack of exercise is a major cause of chronic diseases. *Compr Physiol* 2011; 2: 1143–1211.
8. Shield KD, Parry C and Rehm J. Chronic diseases and conditions related to alcohol use. *Alcohol Res* 2014; 35: 155.
9. Snyder T and Byrd G. The internet of everything. *IEEE Comput Archit Lett* 2017; 50: 8–9.
10. Ravizza S, Huschto T, Adamov A, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat Med* 2019; 25: 57–59.
11. De Choudhury M, Gamon M, Counts S, et al. Predicting depression via social media. In: Proceedings of the 7th international AAAI conference on web and social media, Cambridge, MA, 8–11 July 2013, pp.128–137.
12. Shen JH and Rudzicz F. Detecting anxiety through Reddit. In: Proceedings of the 4th workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality, Vancouver, Canada, 3 August 2017, pp.58–65.
13. Reece AG and Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Sci* 2017; 6: 15.
14. Eichstaedt JC, Smith RJ, Merchant RM, et al. Facebook Language predicts depression in medical records. *Proc Natl Acad Sci USA* 2018; 115: 11203–11208.
15. Kibin. The impact of online shopping on the lifestyle of people in our modern society, http://www.kibin.com/essay-examples/the-impact-of-online-shopping-on-the-lifestyle-of-

people-in-our-modern-society-zeYeyiIq (2021, accessed 23 November 2021)

16. Coppola D. Global number of digital buyers 2014–2021, https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/ (2021, accessed 5 April 2021)

17. Chevalier S. Retail e-commerce sales worldwide from 2014 to 2024, https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/ (2021, accessed 24 November 2021)

18. Cain MM, Sarasohn-Kahn J and Wayne JC. Who are health e-people? A segmentation of online health consumers. In: Fuller B (ed) *Health e-people: The online consumer experience*. Auckland, CA: California HealthCare Foundation, 2000, pp.9–12.

19. Clark D and Weir C. China to surpass US in total retail sales. 2019, https://www.emarketer.com/newsroom/index.php/2019-china-to-surpass-us-in-total-retail-sales/ (2019, accessed 15 January 2021)

20. Blystone D. Understanding the Alibaba business model, https://www.investopedia.com/articles/investing/062315/understanding-alibabas-business-model.asp (2021, accessed 20 January 2021)

21. Sarris J, Thomson R, Hargraves F, et al. Multiple lifestyle factors and depressed mood: A cross-sectional and longitudinal analysis of the UK biobank ($N = 84,860$). *BMC Med* 2020; 18: 354.

22. Reddy PH. Can diabetes be controlled by lifestyle activities? *Curr Res Diabetes Obes J* 2017; 1: 555568.

23. Qin X, Wang S and Hsieh CR. The prevalence of depression and depressive symptoms among adults in China: Estimation based on a national household survey. *China Econ Rev* 2018; 51: 271–282.

24. Que J, Lu L and Shi L. Development and challenges of mental health in China. *Gen Psychiatr* 2019; 32: e100053.

25. Ma RC. Epidemiology of diabetes and diabetic complications in China. *Diabetologia* 2018; 61: 1249–1260.

26. Wang L, Gao P, Zhang M, et al. Prevalence and ethnic pattern of diabetes and prediabetes in China in 2013. *JAMA* 2017; 317: 2515–2523.

27. World Health Organization. Global and regional estimates of prevalence: Depressive disorders. In: *Depression and other common mental disorders: Global health estimates*. Geneva, Switzerland: WHO Document Production Services, 2017, pp.8–9.

28. Wu L. Rate of diabetes in China "explosive", https://www.who.int/china/news/detail/06–04–2016-rate-of-diabetes-in-china-explosive (2016, accessed 15 January 2021)

29. McDonald JH. Multiple logistic regression. In: *Handbook of biological statistics* (3rd edn). Baltimore, MD: Sparky House Publishing, 2014, pp.247–253.

30. Chen H, Cohen P and Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Commun Stat Simul Comput* 2010; 39: 860–864.

31. Ma Y. Distribution of online buyers in China in 2019, by age group, https://www.statista.com/statistics/1172011/china-age-group-distribution-of-online-shoppers/ (2021, accessed 25 November 2021)

32. Coppola D. Distribution of digital buyers in the United States as of February 2020, by age group, https://www.statista.com/statistics/469184/us-digital-buyer-share-age-group/ (2021, accessed 25 November 2021)

33. Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B: Stat Methodol* 1995; 57: 289–300.

34. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.

35. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; 240: 1285–1293.

36. Zisook S, Lesser I, Stewart JW, et al. Effect of age at onset on the course of major depressive disorder. *Am J Psychiatry* 2007; 164: 1539–1546.

37. Elmer T and Stadtfeld C. Depressive symptoms are associated with social isolation in face-to-face interaction networks. *Sci Rep* 2020; 10: 1444.

38. Kim HK and McKenry PC. The relationship between marriage and psychological well-being: A longitudinal analysis. *J Fam Issues* 2020; 23: 885–911.

39. Pedrelli P, Shapero B, Archibald A, et al. Alcohol use and depression during adolescence and young adulthood: A summary and interpretation of mixed findings. *Curr Addict Rep* 2016; 3: 91–97.

40. Gokalp H. Psychosocial aspects of hair loss. In: Kutlubay Z and Serdarogulu S (ed) *Hair and scalp disorders*. London, UK: IntechOpen, 2017, pp.239–252.

41. Simon GE, Ludman EJ, Linde JA, et al. Association between obesity and depression in middle-aged women. *Gen Hosp Psychiatry* 2008; 30: 32–39.

42. West A. Mental health and money—The Anxiety and Depression Association of America (ADAA) weighs in on how financial stress affects your well-being, https://www.badcredit.org/news/adaa-weighs-in-on-how-financial-stress-affects-your-well-being/ (2018, accessed 8 March 2021)

43. Huebschmann AG, Huxley RR, Kohrt WM, et al. Sex differences in the burden of type 2 diabetes and cardiovascular risk across the life course. *Diabetologia* 2019; 62: 1761–1772.

44. Sami W, Ansari T, Butt NS, et al. Effect of diet on type 2 diabetes mellitus: A review. *Int J Health Sci* 2017; 11: 65.

45. Hsu CC, Lee CH, Wahlqvist ML, et al. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. *Diabetes Care* 2012; 35: 2286–2292.

46. Attard SM, Herring AH, Mayer-Davis EJ, et al. Multilevel examination of diabetes in modernising China: What elements of urbanisation are most associated with diabetes? *Diabetologia* 2012; 55: 3182–3192.

47. Lazar MA. How obesity causes diabetes: Not a tall tale. *Science* 2005; 307: 373–375.

48. Cornelis MC, Chiuve SE, Glymour MM, et al. Bachelors, divorcees, and widowers: Does marriage protect men from type 2 diabetes? *PLoS One* 2014; 9: e106720.

49. Trinh NHT, Youn SJ, Sousa J, et al. Using electronic medical records to determine the diagnosis of clinical depression. *Int J Med Inform* 2011; 80: 533–540.

50. Gao WG, Dong YH, Pang ZC, et al. A simple Chinese risk score for undiagnosed diabetes. *Diabet Med* 2010; 27: 274–281.

51. Hariri S, Yoon PW, Qureshi N, et al. Family history of type 2 diabetes: A population-based screening tool for prevention? *Genet Med* 2006; 8: 102–108.

52. Cheung BM and Li C. Diabetes and hypertension: Is there a common metabolic pathway? *Curr Atheroscler Rep* 2012; 14: 160–166.

53. Laudon KC and Traver CG. Social, mobile, and local marketing. In: Wall S (ed) *E-commerce: Business, technology, society* (12th edn). Chicago, IL: RR Donnelley, 2016, pp.460–529.

54. Eaton I and McNett M. Protecting the data: Security and privacy. In: *Data for nurses: Understanding and using data to optimize care delivery in hospitals and health systems* (1st edn). Cambridge, MA: Academic Press, 2019, pp.87–99.

55. O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; 40: 1620–1639.