Reece and Danforth *EPJ Data Science* (2017) 6:15
DOI 10.1140/epjds/s13688-017-0110-z

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Instagram photos reveal predictive markers of depression

Andrew G Reece[1*] and Christopher M Danforth[2,3,4*]

*Correspondence:
reece@g.harvard.edu;
chris.danforth@uvm.edu
[1]Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138, USA
[2]Computational Story Lab, Vermont Advanced Computing Core, University of Vermont, 210 Colchester Ave, Burlington, VT 05405, USA
Full list of author information is available at the end of the article

Abstract

Using Instagram data from 166 individuals, we applied machine learning tools to successfully identify markers of depression. Statistical features were computationally extracted from 43,950 participant Instagram photos, using color analysis, metadata components, and algorithmic face detection. Resulting models outperformed general practitioners' average unassisted diagnostic success rate for depression. These results held even when the analysis was restricted to posts made before depressed individuals were first diagnosed. Human ratings of photo attributes (happy, sad, etc.) were weaker predictors of depression, and were uncorrelated with computationally-generated features. These results suggest new avenues for early screening and detection of mental illness.

Classification: psychological and cognitive sciences; computer science

Keywords: social media; depression; psychology; machine learning; computational social science

1 Introduction

The advent of social media presents a promising new opportunity for early detection and intervention in psychiatric disorders. Predictive screening methods have successfully analyzed online media to detect a number of harmful health conditions [1–11]. All of these studies relied on text analysis, however, and none have yet harnessed the wealth of psychological data encoded in *visual* social media, such as photographs posted to Instagram. In this report, we introduce a methodology for analyzing photographic data from Instagram to predictively screen for depression.

There is good reason to prioritize research into Instagram analysis for health screening. Instagram members currently contribute almost 100 million new posts per day [12], and Instagram's rate of new users joining has recently outpaced Twitter, YouTube, LinkedIn, and even Facebook [13]. A nascent literature on depression and Instagram use has so far either yielded results that are too general or too labor-intensive to be of practical significance for predictive analytics [14, 15]. In particular, Lup et al. [14] only attempted to correlate Instagram usership with depressive symptoms, and Andalibi et al. [15] employed a time-consuming qualitative coding method which the authors acknowledged made it 'impossible to qualitatively analyze' Instagram data at scale (p.4). In our research, we incorporated an ensemble of computational methods from machine learning, image processing,

and other data-scientific disciplines to extract useful psychological indicators from photographic data. Our goal was to successfully identify and predict markers of depression in Instagram users' posted photographs.

Hypothesis 1 *Instagram posts made by individuals diagnosed with depression can be reliably distinguished from posts made by healthy controls, using only measures extracted computationally from posted photos and associated metadata*.

1.1 Photographic markers of depression

Photographs posted to Instagram offer a vast array of features that might be analyzed for psychological insight. The content of photographs can be coded for any number of characteristics: Are there people present? Is the setting in nature or indoors? Is it night or day? Image statistical properties can also be evaluated at a per-pixel level, including values for average color and brightness. Instagram metadata offers additional information: Did the photo receive any comments? How many 'Likes' did it get? Finally, platform activity measures, such as usage and posting frequency, may also yield clues as to an Instagram user's mental state. We incorporated only a narrow subset of possible features into our predictive models, motivated in part by prior research into the relationship between mood and visual preferences.

In studies associating mood, color, and mental health, healthy individuals identified darker, grayer colors with negative mood, and generally preferred brighter, more vivid colors [16–19]. By contrast, depressed individuals were found to prefer darker, grayer colors [17]. In addition, Barrick, Taylor, & Correa [19] found a positive correlation between self-identification with depression and a tendency to perceive one's surroundings as gray or lacking in color. These findings motivated us to include measures of hue, saturation, and brightness in our analysis. We also tracked the use of Instagram filters, which allow users to modify the color and tint of a photograph.

Depression is strongly associated with reduced social activity [20, 21]. As Instagram is used to share personal experiences, it is reasonable to infer that posted photos with people in them may capture aspects of a user's social life. On this premise, we used a face detection algorithm to analyze Instagram posts for the presence and number of human faces in each photograph. We also counted the number of comments and likes each post received as measures of community engagement, and used posting frequency as a metric for user engagement.

1.2 Early screening applications

Hypothesis 1 is a necessary first step, as it addresses an unanswered basic question: Is depression detectable in Instagram posts? On finding support for Hypothesis 1, a natural question arises: Is depression detectable in Instagram posts, *before the date of first diagnosis*? After receiving a depression diagnosis, individuals may come to identify with their diagnosis [22, 23]. Individuals' self-portrayal on social media may then be influenced by this identification. It is possible that a successful predictive model, trained on the entirety of depressed Instagram users' posting histories, might not actually detect depressive signals, per se, but rather purposeful content choices intended to convey a depressive condition. Training a model using only posts made by depressed participants prior to the date of first diagnosis addresses this potential confounding factor.

Hypothesis 2 *Instagram posts made by depressed individuals prior to the date of first clinical diagnosis can be reliably distinguished from posts made by healthy controls*.

If support is found for Hypothesis 2, this would not only demonstrate a methodological advance for researchers, but also serve as a proof-of-concept for future healthcare applications. As such, we benchmarked the accuracy of our model against the ability of general practitioners to correctly diagnose depression as shown in a meta-analysis by Mitchell, Vaze, and Rao [24]. The authors analyzed 118 studies that evaluated general practitioners' abilities to correctly diagnose depression in their patients, without assistance from scales, questionnaires, or other measurement instruments. Out of 50,371 patient outcomes included across the pooled studies, 21.9% were actually depressed, as evaluated separately by psychiatrists or validated interview-based measures conducted by researchers. General practitioners were able to correctly rule out depression in non-depressed patients 81% of the time, but only diagnosed depressed patients correctly 42% of the time. We refer to these meta-analysis findings [24] as a comparison point to evaluate the usefulness of our models.

A major strength of our proposed models is that their features are generated using entirely computational means - pixel analysis, face detection, and metadata parsing - which can be done at scale, without additional human input. It seems natural to wonder whether these machine-extracted features pick up on similar signals that humans might use to identify mood and psychological condition, or whether they attend to wholly different information. A computer may be able to analyze the average saturation value of a million pixels, but can it pick out a happy selfie from a sad one? Understanding whether machine learning and human opinion are sensitive to the same indicators of depression may be valuable information for future research and applications. Furthermore, insight into these issues may help to frame our results in the larger discussion around human versus machine learning, which occupies a central role in the contemporary academic landscape.

To address these questions, we solicited human assessments of the Instagram photographs we collected. We asked new participants to evaluate photos on four simple metrics: happiness, sadness, interestingness, and likability. These ratings categories were intended to capture human impressions that were both intuitive and quantifiable, and which had some relationship to established depression indicators. DSM-IV [20] criteria for Major Depressive Disorder includes feeling sad as a primary criterion, so sadness (and its anti-correlate, happiness) seemed obvious candidates as ratings categories. Epstein et al. [25] found depressed individuals 'had difficulty reconciling a self-image as an 'outgoing likeable person'', which prompted likability as an informative metric. We hypothesized that human raters should find photographs posted by depressed individuals to be sadder, less happy, and less likable, on average. Finally, we considered interestingness as a novel factor, without a clear directional hypothesis.

Hypothesis 3a *Human ratings of Instagram posts on common semantic categories can distinguish between posts made by depressed and healthy individuals*.

Hypothesis 3b *Human ratings are positively correlated with computationally-extracted features*.

If human and machine[a] predictors show positive correlation, we can infer that each set of features tracks similar signals of depression. In this case, the strength of the human model simply suggests whether it is better or worse than the machine model. On the other hand, if machine and human features show little or no correlation, then regardless of human model performance, we would know that the machine features are capable of screening for depression, but use different information signals than what are captured by the affective ratings categories.

2 Method

2.1 Data Collection

Data collection was crowdsourced using Amazon's Mechanical Turk (MTurk) crowdwork platform. Separate surveys were created for depressed and healthy individuals. In the depressed survey, participants were invited to complete a survey that involved passing a series of inclusion criteria, responding to a standardized clinical depression survey, answering questions related to demographics and history of depression, and sharing social media history. We used the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire to screen participant depression levels [26]. CES-D assessment quality has been demonstrated as on-par with other depression inventories, including the Beck Depression Inventory and the Kellner Symptom Questionnaire [27, 28]. Healthy participants were screened to ensure no history of depression and active Instagram use. See Additional file 1 for actual survey text.

Qualified participants were asked to share their Instagram usernames and history. An app embedded in the survey allowed participants to securely log into their Instagram accounts and agree to share their data.[b] Upon securing consent, we made a one-time collection of participants' entire Instagram posting history. In total we collected 43,950 photographs from 166 Instagram users, 71 of whom had a history of depression.

We asked a different set of MTurk crowdworkers to rate the Instagram photographs collected. This new task asked participants to rate a random selection of 20 photos from the data we collected. Raters were asked to judge how interesting, likable, happy, and sad each photo seemed, on a continuous 0-5 scale. Each photo was rated by at least three different raters, and ratings were averaged across raters. Raters were not informed that photos were from Instagram, nor were they given any information about the study participants who provided the photos, including mental health status. Each ratings category showed good inter-rater agreement.

Only a subset of participant Instagram photos were rated (N = 13,184). We limited ratings data to a subset because this task was time-consuming for crowdworkers, and so proved a costly form of data collection. For the depressed sample, ratings were only made for photos posted within a year in either direction of the date of first depression diagnosis. Within this subset, for each user the nearest 100 posts prior to the diagnosis date were rated. For the control population, the most recent 100 photos from each user's date of participation in this study were rated.

2.2 Participant safety and privacy

Data privacy was a concern for this study. Strict anonymity was nearly impossible to guarantee to participants, given that usernames and personal photographs posted to Instagram often contain identifiable features. We made sure participants were informed of the risks

Figure 1 Comparison of HSV values. Right photograph has higher Hue (bluer), lower Saturation (grayer), and lower Brightness (darker) than left photograph. Instagram photos posted by depressed individuals had HSV values shifted towards those in the right photograph, compared with photos posted by healthy individuals.

of being personally identified, and assured them that no data with personal identifiers, including usernames, would be made public or published in any format.

2.3 Improving data quality

We employed several quality assurance measures in our data collection process to reduce noisy and unreliable data. Our surveys were only visible to MTurk crowdworkers who had completed at least 100 previous tasks with a minimum 95% approval rating; MTurk workers with this level of experience and approval rating have been found to provide reliable, valid survey responses [29]. We also restricted access to only American IP addresses, as MTurk data collected from outside the United States are generally of poorer quality [30]. All participants were only permitted to take the survey once.

We excluded participants who had successfully completed our survey, but who had a lifetime total of fewer than five Instagram posts. We also excluded participants with CES-D scores of 22 or higher. Studies have indicated that a CES-D score of 22 represents an optimal cutoff for identifying clinically relevant depression across a range of age groups and circumstances [31, 32].

2.4 Feature extraction

Several different types of information were extracted from the collected Instagram data. We used total posts per user, per day, as a measure of user activity. We gauged community reaction by counting the number of comments and 'likes' each posted photograph received. Face detection software was used to determine whether or not a photograph contained a human face, as well as count the total number of faces in each photo, as a proxy measure for participants' social activity levels. Pixel-level averages were computed for Hue, Saturation, and Value (HSV), three color properties commonly used in image analysis. Hue describes an image's coloring on the light spectrum (ranging from red to blue/purple). Lower hue values indicate more red, and higher hue values indicate more blue. Saturation refers to the vividness of an image. Low saturation makes an image appear grey and faded. Value refers to image brightness. Lower brightness scores indicate a darker image. See Figure 1 for a comparison of high and low HSV values. We also checked metadata to assess whether an Instagram-provided filter was applied to alter the appearance of a photograph. Collectively, these measures served as the feature set in our primary

model. For the separate model fit on ratings data, we used only the four ratings categories (happy, sad, likable, interesting) as predictors.

2.5 Units of observation

In determining the best time span for this analysis, we encountered a difficult question: When and for how long does depression occur? A diagnosis of depression does not indicate the persistence of a depressive state for every moment of every day, and to conduct analysis using an individual's entire posting history as a single unit of observation is therefore rather specious. At the other extreme, to take each individual photograph as units of observation runs the risk of being too granular. De Choudhury et al. [2] looked at all of a given user's posts in a single day, and aggregated those data into per-person, per-day units of observation. We adopted this precedent of 'user-days' as a unit of analysis.[c]

2.6 Statistical framework

We used Bayesian logistic regression with uninformative priors to determine the strength of individual predictors. Two separate models were trained. The All-data model used all collected data to address Hypothesis 1. The Pre-diagnosis model used all data collected from healthy participants, but only pre-diagnosis data from depressed participants, to address Hypothesis 2. We also fit an 'intercept-only' model, in which all predictors are zero-weighted to simulate a model under a null hypothesis. Bayes factors were used to assess model fit. Details on Bayesian estimation, model optimization and selection, and diagnostic checks are available in Additional file 1.

We also employed a suite of supervised machine learning algorithms to estimate the predictive capacity of our models. We report prediction results only from the best-performing algorithm, a 100-tree Random Forests classifier. As an informal benchmark for comparison, we present general practitioners' unassisted diagnostic accuracy as reported in Mitchell, Vaze, and Rao [24].[d]

In evaluating binary classification accuracy, a simple proportion of correct classifications is often inappropriate. In cases where data exhibit a class imbalance, i.e. more healthy than depressed observations (or vice-versa), reporting naive accuracy can be misleading. (A classification accuracy of 95% seems excellent until it is revealed that 95% of the data modeled belong to a single class.) Additionally, naive accuracy scores are opaque to the specific strengths and weaknesses of a binary classifier. Instead, we report precision, recall, specificity, negative predictive value, and F1 scores for fuller context. Definitions for these terms are as follows:

Precision	TP/(TP + FP)
Recall	TP/(TP + FN)
Specificity	TN/(TN + FP)
Negative Predictive Value	TN/(TN + FN)
F1	$2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$
TP = True Positive FP = False Positive TN = True Negative FN = False Negative	

3 Results

Both All-data and Pre-diagnosis models were decisively superior to a null model (K_{all} = 157.5; K_{pre} = 149.8), see page 4 of the Additional file 1 for a description of K. All-data predictors were significant with 99% probability. Pre-diagnosis and All-data confidence levels were largely identical, with two exceptions: Pre-diagnosis Brightness decreased to 90%

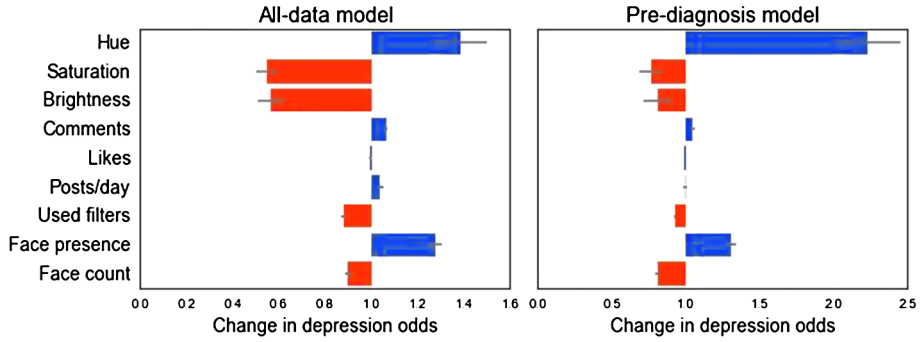

Figure 2 Magnitude and direction of regression coefficients in All-data (*N* = 24,713) and Pre-diagnosis (*N* = 18,513) models. X-axis values represent the adjustment in odds of an observation belonging to the target class, per unit increase of each predictive variable. Odds were generated by exponentiating logistic regression log-odds coefficients.

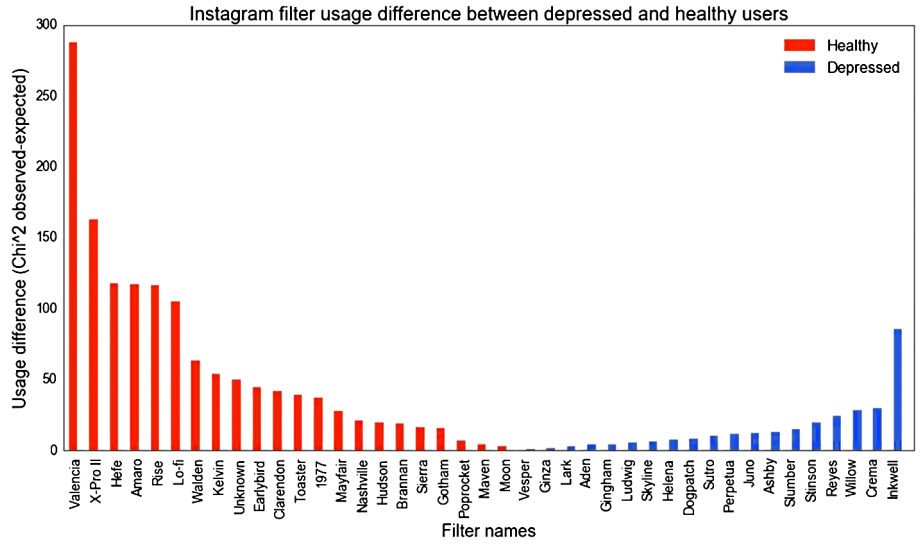

Figure 3 Instagram filter usage among depressed and healthy participants. Bars indicate difference between observed and expected usage frequencies, based on a Chi-squared analysis of independence. Blue bars indicate disproportionate use of a filter by depressed compared to healthy participants, orange bars indicate the reverse. All-data results are displayed, see Additional file 1 for Pre-diagnosis plot.

confidence, and Pre-diagnosis posting frequency dropped to 30% confidence, suggesting a null predictive value in the latter case.

Increased hue, along with decreased brightness and saturation, predicted target class observations. This means that photos posted by depressed individuals tended to be bluer, darker, and grayer (see Figure 1). The more comments Instagram posts received, the more likely they were posted by depressed participants, but the opposite was true for likes received. In the All-data model, higher posting frequency was also associated with depression. Depressed participants were more likely to post photos with faces, but had a lower average face count per photograph than healthy participants. Finally, depressed participants were less likely to apply Instagram filters to their posted photos. Figure 2 shows the magnitude and direction of regression coefficients for both models.

Table 1 Comparison of accuracy metrics for All-data and Pre-diagnosis model predictions

	Mitchell et al. μ	All-data $\mu(\sigma)$	Pre-diagnosis $\mu(\sigma)$
Recall	0.510	0.697 (0.008)	0.318 (0.012)
Specificity	0.813	0.478 (0.012)	0.833 (0.010)
Precision	0.42	0.604 (0.009)	0.541 (0.009)
Negative Predictive Value	0.858	0.579 (0.008)	0.665 (0.006)
F1	0.461	0.647 (0.003)	0.401 (0.008)

General practitioners' diagnostic accuracy from (Mitchell et al. [24]) is included for comparison. See see Additional file 1 for definitions of accuracy metrics.

A closer look at filter usage in depressed versus healthy participants provided additional texture. Instagram filters were used differently by target and control groups ($\chi^2_{\text{all}} = 907.84$, $p = 9.17 \times 10^{-164}$; $\chi^2_{\text{pre}} = 813.80$, $p = 2.87 \times 10^{-144}$). In particular, depressed participants were less likely than healthy controls to use any filters at all. When depressed participants did employ filters, they most disproportionately favored the 'Inkwell' filter, which converts color photographs to black-and-white images (see Figure 3). Conversely, healthy participants most disproportionately favored the Valencia filter, which lightens the tint of photos. Examples of filtered photographs are provided in Additional file 1.

Our best All-data machine learning classifier, averaged over five randomized iterations, improved over Mitchell et al. [24] general practitioner accuracy on most metrics (see Table 1). Compared with Mitchell et al. [24] results, the All-data model was less conservative (lower specificity) but better able to positively identify target class observations (higher recall). Given 100 observations, our model correctly identified 70% of all target class cases ($n = 37$), with a relatively low number of false alarms ($n = 23$) and misses ($n = 17$).

Pre-diagnosis predictions showed improvement over the Mitchell et al. [24] benchmark on precision and specificity. The Pre-diagnosis model found only about a third of actual target class observations, but it was correct most of the time when it did predict a target class label. By comparison, although Mitchell et al. [24] general practitioners discovered more true cases of depression, they were more likely than not to misdiagnose healthy subjects as depressed.

Out of the four predictors used in the human ratings model (happiness, sadness, likability, interestingness), only the sadness and happiness ratings were significant predictors of depression. Depressed participants' photos were more likely to be sadder and less happy than those of healthy participants. Ratings assessments generally showed strong patterns of correlation with one another, but exhibited extremely low correlation with computational features. The modest positive correlation of human-rated happiness with the presence and number of faces in a photograph was the only exception to this trend. Correlation matrices for all models are available in Additional file 1.

4 Discussion

The present study employed computational machine learning techniques to screen for depression using photographs posted to Instagram. Our results supported Hypothesis 1, that markers of depression are observable in Instagram user behavior, and Hypothesis 2, that these depressive signals are detectable in posts made even before the date of first diagnosis. Human ratings proved capable of distinguishing between Instagram posts made by depressed and healthy individuals (Hypothesis 3a), but showed little or no correlation with most computational features (Hypothesis 3b). Our findings establish that visual social

media data are amenable to analysis of affect using scalable, computational methods. One avenue for future research might integrate textual analysis of Instagram posts' comments, captions, and tags. Considering the early success of textual analysis in detecting various health and psychological signals on social media [5, 33, 34], the modeling of textual and visual features together could well prove superior to either medium on its own.

Our model showed considerable improvement over the ability of unassisted general practitioners to correctly diagnose depression. On average, more than half of general practitioners' depression diagnoses were false positives [24]. By comparison, the majority of both All-data and Pre-diagnosis depression classifications were correct. As false diagnoses are costly for both healthcare programs and individuals, this improvement is noteworthy. Health care providers may be able to improve quality of care and better identify individuals in need of treatment based on the simple, low-cost methods outlined in this report. Given that mental health services are unavailable or underfunded in many countries [35], this computational approach, requiring only patients' digital consent to share their social media histories, may open avenues to care which are currently difficult or impossible to provide.

On the other hand, our Pre-diagnosis prediction engine was rather conservative, and tended to classify most observations as healthy. There is good reason to believe, however, that the Pre-diagnosis prediction accuracy observed represents a lower bound on performance. Ideally, we would have used the All-data classifier to evaluate the Pre-diagnosis data, as that model was trained on a much larger dataset. The fact that the Pre-diagnosis data was a subset of the full dataset meant that applying the All-data model to Pre-diagnosis observations would have artificially inflated accuracy, due to information leakage between training and test data. Instead, we trained a new classifier for Pre-diagnosis, using training and test partitions contained within the Pre-diagnosis data, which left the Pre-diagnosis model with considerably fewer data points to train on. As a result, it is likely that Pre-diagnosis accuracy scores understate the technique's true capacity.

Regarding the strength of specific predictive features, some results match common perceptions regarding the effects of depression on behavior. Photos posted to Instagram by depressed individuals were more likely to be bluer, grayer, and darker, and receive fewer likes. Depressed Instagram users in our sample had an outsized preference for filtering out all color from posted photos, and showed an aversion to artificially lightening photos, compared to non-depressed controls. These results matched well with the literature linking depression and a preference for darker, bluer, and monochromatic colors [16–19]. Depressed users were more likely to post photos with faces, but they tended to post fewer faces per photo. This finding may be an oblique indicator that depressed users interact in smaller social settings, or at least choose only to share experiences of this sort on social media. This would be in accordance with previous findings that reduced social interactivity is an indicator of depression [5, 20, 21].

Other, seemingly obvious, relationships failed to emerge. For example, when people rated a photograph as sad, that impression was unrelated to how blue, dark, or gray that photo was. Both 'sad' and 'blue, dark, and gray' were strong predictors of depression, however, and semantically these descriptions seem like they should match well with one another, as well as link to depression. These divergences may serve as the basis for a number of future research inquiries into the relationship between depressive behavior and common perceptions of depression.

A general limitation to these findings concerns the non-specific use of the term 'depression' in the data collection process. We acknowledge that depression describes a general clinical status, and is frequently comorbid with other conditions. It is possible that a specific diagnostic class is responsible for driving the observed results, and future research should fine-tune questionnaires to acquire specific diagnostic information. Additionally, it is possible that our results are in some way specific to individuals who received clinical diagnoses. Current perspectives on depression treatment indicate that people who are 'well-informed and psychologically minded, experience typical symptoms of depression and little stigma, and have confidence in the effectiveness of treatment, few concerns about side effects, adequate social support, and high self-efficacy' seek out mental health services [25]. The intersection of these qualities with typical Instagram user demographics suggests caution in making broad inferences, based on our findings.

As these methods provide a tool for inferring personal information about individuals, two points of caution should be considered. First, data privacy and ethical research practices are of particular concern, given recent admissions that individuals' social media data were experimentally manipulated or exposed without permission [36, 37]. It is perhaps reflective of a current general skepticism towards social media research that, of the 509 individuals who began our survey, 221 (43%) refused to share their Instagram data, even after we provided numerous privacy guarantees. Future research should prioritize establishing confidence among experimental participants that their data will remain secure and private. Second, data trends often change over time, leading socio-technical models of this sort to degrade without frequent calibration [38]. The findings reported here should not be taken as enduring facts, but rather as promising leads upon which to build and refine subsequent models.

Paired with a commensurate focus on upholding data privacy and ethical analytics, the present work may serve as a blueprint for effective mental health screening in an increasingly digitalized society. More generally, these findings support the notion that major changes in individual psychology are transmitted in social media use, and can be identified via computational methods.

Additional material

Additional file 1: Supplementary materials. (pdf)

Acknowledgements
The authors thank K Lix for conversations and manuscript review.

Funding
CMD acknowledges funding from the National Science Foundation under Grant No. IIS-1447634. AGR acknowledges support from the Sackler Scholar Programme in Psychobiology.

Availability of data and materials
Code associated with this study is available publicly on the github page of AGR:
https://github.com/andrewreece/predicthealth.

Ethics approval and consent to participate
This study was reviewed and approved by the Harvard University Institutional Review Board, approval #15-2529 and by the University of Vermont Institutional Review Board, approval #CHRMS-16-135.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
N/A

Authors' contributions
AGR and CMD designed the study. AGR performed the study and analyzed results. AGR and CMD authored the manuscript.

Author details
[1]Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138, USA. [2]Computational Story Lab, Vermont Advanced Computing Core, University of Vermont, 210 Colchester Ave, Burlington, VT 05405, USA.
[3]Department of Mathematics and Statistics, University of Vermont, 210 Colchester Ave, Burlington, VT 05405, USA.
[4]Vermont Complex Systems Center, University of Vermont, 210 Colchester Ave, Burlington, VT 05405, USA.

Endnotes
[a] The term 'machine' (e.g. 'machine predictors', 'machine model') is used as shorthand for the computational feature extraction process we employed. Significant human biases informed this process, however, as the initial selection of features for extraction involved entirely human decision-making.
[b] Data collection source code is available on Github, see Additional file 1.
[c] Occasionally, when reporting results we refer to 'observations' as 'participants', e.g. 'depressed participants received fewer likes'. It would be more correct to use the phrase 'photographic data aggregated by participant-user-days' instead of 'participants'. We chose to sacrifice a degree of technical correctness for the sake of clarity.
[d] Comparing point estimates of accuracy metrics is not a statistically robust means of model comparison. However, we felt it was more meaningful to frame our findings in a realistic context, rather than to benchmark against a naive statistical model that simply predicted the majority class for all observations.

Publisher's Note

Received: 28 March 2017 Accepted: 22 June 2017 Published online: 08 August 2017

References
1. Moreno M, Christakis D, Egan K, Brockman L, Becker T (2012) Associations between displayed alcohol references on Facebook and problem drinking among college students. Arch Pediatr Adolesc Med 166(2):157-163. doi:10.1001/archpediatrics.2011.180
2. De Choudhury M, Counts S, Horvitz E (2013) Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 3267-3276. doi:10.1145/2470654.2466447
3. De Choudhury M, Counts S, Horvitz EJ, Hoff A (2014) Characterizing and predicting postpartum depression from shared Facebook data. In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing, ACM, New York, pp 626-638. doi:10.1145/2531602.2531675
4. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media
5. Katikalapudi R, Chellappan S, Montgomery F, Wunsch D, Lutzen K (2012) Associating Internet usage with depressive behavior among college students. IEEE Technol Soc Mag 31(4):73-80. doi:10.1109/MTS.2012.2225462
6. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, Becker T (2011) Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety 28(6):447-455. doi:10.1002/da.20805
7. Coppersmith G, Harman C, Dredze M (2014) Measuring post traumatic stress disorder in Twitter. In: Eighth international AAAI conference on weblogs and social media
8. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M (2016) Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI conference on human factors in computing systems. ACM, New York, pp 2098-2110. doi:10.1145/2858036.2858207
9. Christakis NA, Fowler JH (2010) Social network sensors for early detection of contagious outbreaks. PLoS ONE 5(9):e12948. doi:10.1371/journal.pone.0012948
10. Schmidt CW (2012) Trending now: using social media to predict and track disease outbreaks. Environ Health Perspect 120(1):a30-a33. doi:10.1289/ehp.120-a30
11. Paparrizos J, White RW, Horvitz E (2016) Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results J Oncol Pract 12(8):737-744. doi:10.1200/JOP.2015.010504
12. Instagram (2016) Instagram press release. Available at https://www.instagram.com/press/. Accessed July 26, 2016
13. Chaffey D (2016) Global social media research summary 2016. Available at bit.ly/1WRviEl. Accessed July 19, 2016
14. Lup K, Trub L, Rosenthal L (2015) Instagram #Instasad?: exploring associations among Instagram use, depressive symptoms, negative social comparison, and strangers followed. Cyberpsychol Behav 18(5):247-252. doi:10.1089/cyber.2014.0560
15. Andalibi N, Ozturk P, Forte A (2015) Depression-related imagery on Instagram. In: Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing, ACM, New York, pp 231-234. doi:10.1145/2685553.2699014
16. Boyatzis CJ, Varghese R (1994) Children's emotional associations with colors. J Genet Psychol 155(1):77-85
17. Carruthers HR, Morris J, Tarrier N, Whorwell PJ (2010) The Manchester Color Wheel: development of a novel way of identifying color choice and its validation in healthy, anxious and depressed individuals. BMC Med Res Methodol 10:12. doi:10.1186/1471-2288-10-12

18. Hemphill M (1996) A note on adults' color-emotion associations. J Genet Psychol 157(3):275-280
19. Barrick CB, Taylor D, Correa EI (2002) Color sensitivity and mood disorders: biology or metaphor? J Affect Disord 68(1):67-71. doi:10.1016/S0165-0327(00)00358-X
20. American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders, 4th edn. doi:10.1176/appi.books.9780890423349
21. Bruce ML, Hoff RA (1994) Social and physical health risk factors for first-onset major depressive disorder in a community sample. Soc Psychiatry Psychiatr Epidemiol 29(4):165-171. doi:10.1007/BF00802013
22. Cornford CS, Hill A, Reilly J (2007) How patients with depressive symptoms view their condition: a qualitative study. Fam Pract 24(4):358-364. doi:10.1093/fampra/cmm032
23. Karp DA (1994) Living with depression: illness and identity turning points. Qual Health Res 4(1):6-30. doi:10.1177/104973239400400102
24. Mitchell AJ, Vaze A, Rao S (2009) Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 374(9690):609-619. doi:10.1016/S0140-6736(09)60879-5
25. Epstein RM, Duberstein PR, Feldman MD, Rochlen AB, Bell RA, Kravitz RL et al (2010) 'I didn't know what was wrong:' how people with undiagnosed depression recognize, name and explain their distress. J Gen Intern Med 25(9):954-961. doi:10.1007/s11606-010-1367-0
26. Radloff LS (1977) The CES-D scale: a self-report depression scale for research in the general population. Appl Psych Manage 1(3):385-401. doi:10.1177/014662167700100306
27. Fountoulakis KN, Bech P, Panagiotidis P, Siamouli M, Kantartzis S, Papadopoulou A et al (2007) Comparison of depressive indices: reliability, validity, relationship to anxiety and personality and the role of age and life events. J Affect Disord 97(1-3):187-195. doi:10.1016/j.jad.2006.06.015
28. Zich JM, Attkisson CC, Greenfield TK (1990) Screening for depression in primary care clinics: the CES-D and the BDI. Int J Psychiatry Med 20(3):259-277. doi:10.2190/LYKR-7VHP-YJEM-MKM2
29. Peer E, Vosgerau J, Acquisti A (2013) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behav Res Methods 46(4):1023-1031. doi:10.3758/s13428-013-0434-y
30. Litman L, Robinson J, Rosenzweig C (2014) The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. Behav Res Methods 47(2):519-528. doi:10.3758/s13428-014-0483-x
31. Cuijpers P, Boluijt B, van Straten A (2007) Screening of depression in adolescents through the Internet. Eur Child Adolesc Psychiatry 17(1):32-38. doi:10.1007/s00787-007-0631-2
32. Haringsma R, Engels GI, Beekman ATF, Spinhoven P (2004) The criterion validity of the center for epidemiological studies depression scale (CES-D) in a sample of self-referred elders with depressive symptomatology. Int J Geriatr Psychiatry 19(6):558-563. doi:10.1002/gps.1130
33. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6(12):e26752. doi:10.1371/journal.pone.0026752
34. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ (2016) Forecasting the onset and course of mental illness with Twitter data. arXiv:1608.07740
35. Detels R (2009) The scope and concerns of public health. Oxford University Press, London
36. Fiske ST, Hauser RM (2014) Protecting human research participants in the age of big data. Proc Natl Acad Sci USA 111(38):13675-13676. doi:10.1073/pnas.1414626111
37. Lumb D (2016) Scientists release personal data for 70,000 OkCupid profiles. Available at engt.co/2b4NnQ0. Accessed August 7, 2016
38. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203-1205. doi:10.1126/science.1248506
39. Gigerenzer G (2004) Mindless statistics. J Socio-Econ 33(5):587-606. doi:10.1016/j.socec.2004.09.03
40. Hubbard R, Lindsay RM (2008) Why p-values are not a useful measure of evidence in statistical significance testing. Theory Psychol 18(1):69-88. doi:10.1177/0959354307086923
41. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ (2015) The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 23(1):103-123. doi:10.3758/s13423-015-0947-8
42. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. Am Stat 70(2):129-133. doi:10.1080/00031305.2016.1154108
43. Martin A, Quinn K, Park JH (2011) MCMCpack: Markov chain Monte Carlo in R. J Stat Softw 42(9):1-21
44. Link WA, Eaton MJ (2012) On thinning of chains in MCMC. Methods Ecol Evol 3(1):112-115. doi:10.1111/j.2041-210X.2011.00131.x
45. Christensen R, Johnson W, Branscum A, Hanson TE (2011) Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC Press, Boca Raton
46. Jeffries H (1961) Theory of probability. Clarendon, Oxford
47. Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Vol. 196. Federal Reserve Bank of Minneapolis, Research Department, Minneapolis
48. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457-472
49. Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis (Vol. 2). CRC Press, Boca Raton