CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING
Volume 21, Number 4, 2018

DOI: 10.1089/cyber.2017.0384

REVIEW

Predicting Individual Characteristics from Digital Traces on Social Media: A Meta-Analysis

Michele Settanni, Danny Azucar, and Davide Marengo

Abstract

The increasing utilization of social media provides a vast and new source of user-generated ecological data (digital traces), which can be automatically collected for research purposes. The availability of these data sets, combined with the convergence between social and computer sciences, has led researchers to develop automated methods to extract digital traces from social media and use them to predict individual psychological characteristics and behaviors. In this article, we reviewed the literature on this topic and conducted a series of meta-analyses to determine the strength of associations between digital traces and specific individual characteristics; personality, psychological well-being, and intelligence. Potential moderator effects were analyzed with respect to type of social media platform, type of digital traces examined, and study quality. Our findings indicate that digital traces from social media can be studied to assess and predict theoretically distant psychosocial characteristics with remarkable accuracy. Analysis of moderators indicated that the collection of specific types of information (i.e., user demographics), and the inclusion of different types of digital traces, could help improve the accuracy of predictions.

Keywords: social media, digital traces, psychosocial characteristics, psychological assessment, data mining, predictive modeling

Introduction

Emergence of social media

THE RECENT YEARS have seen a major evolution in how people interact online,[1] and the growth of social media has yielded great sources of online interpersonal communication, with users expressing themselves in naturalistic settings about everyday topics and events.[2–6] This ever-increasing utilization of social media provides a vast and new source of user-generated ecological data with connections to offline personal characteristics, attitudes, and behaviors.[5–11] Digitally mediated behaviors on social media are recorded and can be collected and analyzed by researchers from diverse disciplines. More specifically, due to the popularity of social media, psychologists have begun studying the relationships between psychosocial characteristics and digitally mediated human behaviors, or ''*digital traces*.''[6,12–14]

The terms ''digital traces,'' ''digital footprints,'' and ''digital records'' are used interchangeably; throughout this article we use ''digital traces'' for consistency. Defined as information generated by users on their social media profiles, digital traces consist of personal information about age, gender, sexual orientation, and location, as well as activity information, including network size, shared text, pictures, and videos.[15] The availability of large data sets from social media, fostered by the convergence between social and computer sciences, allows researchers to not only seek to *gain insights* from studying human behaviors on social media but also to *predict* psychological characteristics and behaviors based on automated data mining and the analysis of digital traces.

Studies using automated approaches mostly aim at developing models to predict individual characteristics using the information available on social media profiles (e.g., predicting personality using data referring to activity statistics, language use, and pictures posted on social media).[6,14] These type of researches are mainly data-driven, using features extracted from digital traces to predict psychological characteristics without referring to specific *a priori* theories or hypotheses.[6,16] Studies in this field focus on association rather than causation, without providing theoretical explanation of the relationship between predictors and outcomes.[17] The present review focuses on this type of researches.

Department of Psychology, University of Turin, Turin, Italy.

Predicting individual characteristics via automated analysis of digital traces

Studies focusing on the prediction of psychosocial and behavioral characteristics based on digital traces from social media generally use a common methodology, consisting of the following steps: (1) users are contacted and asked to complete self-report questionnaires assessing the characteristics of interest, and provide complete or limited access to their digital traces on social media, (2) digital traces are collected and analyzed using automated approaches to extract sets of profile attributes, or *features (*e.g., *activity statistics*, such as number of friends, and status updates; *linguistic features*, such as frequency of words in predefined categories in posts), and (3) the predictive power of these features is examined over participants' individual characteristics as assessed via self-reports, using a varied set of predictive methods, ranging from univariate linear regression modeling to classification via machine learning.

One of the earliest projects using this approach is the MyPersonality project,[13] which has cultivated a data set consisting of self-report data for a wide range of psychosocial characteristics (e.g., personality, satisfaction with life, substance use), and digital traces of nearly 3 million Facebook users. Many researchers have used this data set to conduct automatic coding of user profiles and assess or predict distinguishing features of user personality and well-being.[5,6,9,10,13,18–22] Furthermore, scholars have demonstrated the feasibility of predicting many psychosocial characteristics from features extracted from a variety of digital traces (e.g., user demographics,[20] activity statistics,[8,18] linguistic features,[6] and features extracted from pictures[23]), social media platforms (e.g., Facebook,[24] Twitter,[25] and Sina Weibo[2]), and by using different analytical approaches (e.g., use of a single type of digital trace[8] vs. multiple sources of digital traces[20]).

Aims

The aim of the current study is to conduct a series of meta-analyses to determine the mean effect size of associations between digital traces from social media and specific individual characteristics. Meta-analyses were conducted on characteristics investigated by at least three studies, namely personality, psychological well-being, and intelligence. Given the expected presence of effect size heterogeneity among studies, potential moderator effects were analyzed with respect to type of social media platform (public vs. private), type of digital traces, and study quality.

Materials and Methods

Search strategy and inclusion criteria

An initial data set of 1,677 articles was identified in July 2016 by submitting a search query to Scopus, ISI Web of Science, PubMed, and ProQuest databases. The query searched keywords in the ''title,'' ''abstract,'' and ''keyword heading'' fields. The following keywords and stems were used in separate and combined searches:

psych, behavior, personality, health, well-being, risk, depression, quality-of-life, life satisfaction, risk-behavior, substance, abuse, psychological-assessment, cyberpsychology, emotional-well-being, mental-health, gender, age,* in conjunction with *Myspace, Facebook, Instagram, Twitter, Youtube, Photobucket, Linkedin, social network, Reddit, social media, Snapchat, Periscope, social-networking, status-updates, mypersonality, machine-learning, data-mining, text-analysis, language-processing, closed-vocabulary, closed-dictionary, liwc, open-vocabulary, open-dictionary, support-vector-machines, text-mining, topic-modeling, dictionary, latent-dirichlet-allocation, differential-language-analysis, digital-footprint, differential-language, computational-social-science, content-analysis, linguistic-studies*

After duplicates were removed, a set of 1,241 articles was screened for the following inclusion criteria—(1) studies must focus on human behavioral or psychological characteristics, (2) studies must focus on *individual* human behavioral or psychological characteristics, (3) studies must be at least partially quantitative in nature, (4) studies must analyze digital traces of human behavior, and (5) studies must include a valid self-report measure to assess individual characteristics. A total of 1,203 articles were excluded on inspection of their abstracts, and full-text assessment for eligibility was conducted for 38 articles. This screening process resulted in the initial selection of 25 articles for inclusion in our analysis.

We then identified an additional 34 articles through a review of the ''citations'' from the 25 originally selected articles; of these, 13 were selected for inclusion in our review based on the aforementioned inclusion criteria. This resulted in a final set of 38 articles selected for the review. The article selection process is depicted in Figure 1.

Research coding

Coding of psychological and behavioral characteristics. Investigated characteristics varied across studies. We identified three general psychological characteristics, which were investigated at least by three studies: personality traits (Big 5 and Dark Triad), psychological well-being (depression, anxiety and stress, life satisfaction), and intelligence. Other characteristics were present in less than three studies: personal values, coping strategies, substance use, and self-monitoring skills (Table 1).

Coding of digital traces. Studies varied considerably in terms of digital traces analyzed. We distinguished between the following types of digital traces: (1) user demographics (e.g., gender, age, and location), (2) user activity statistics (e.g., number of posts, number of friends, number of likes, comments, and mentions), (3) language (e.g., tweets, status updates, and comments), (4) Facebook likes (i.e., expression of interest in Facebook pages about events, persons, locations, and products), and (5) pictures (e.g., profile pictures and Instagram photos).

Coding of moderators. We considered seven potential moderators that were dichotomously coded. (1) Type of social media platform (private vs. public), (2) use of user demographics (yes vs. no), (3) use of activity statistics (yes vs. no), (4) use of language-based features (yes vs. no), (5) use of pictures (yes vs. no), (6) use of multiple versus single types of digital traces (e.g., language vs. language+pictures), and (7) study quality. Concerning the distinction between types of social media platform, we chose to group social media based on their default privacy settings, distinguishing between public (platforms that make posts and updates public

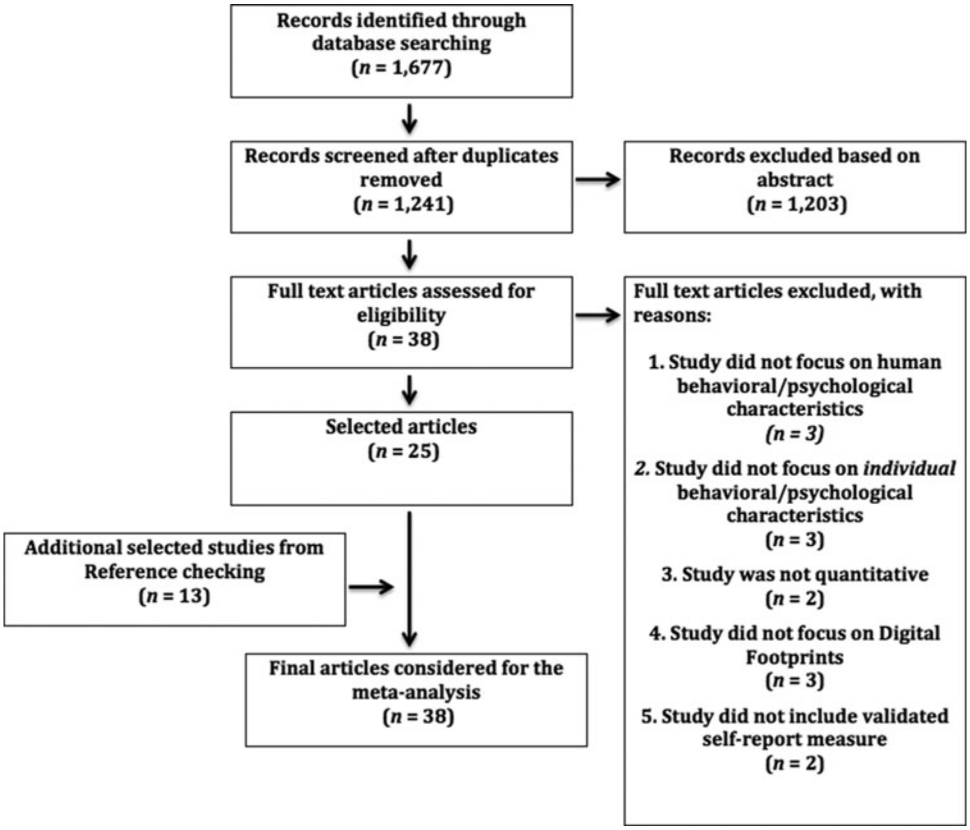

FIG. 1. Flowchart of article selection.

by default, i.e., Twitter, Sina Weibo, Reddit, and Instagram) and private (platforms in which user posts are visible only by users' friends, i.e., Facebook).

Concerning study quality, given that the heterogeneity of the research areas in which analyzed studies were conducted makes it impossible to define a methodological standard, study quality was assessed using the quality of the source the study was published in. Articles were categorized into top, middle, and low tiers using the quartile that sources belong to in the 2016 Scopus CiteScore; ranking quartile 1 as top tier (high quality), quartile 2 as middle tier (medium quality), and quartile 3 or 4 and nonindexed studies as low tier (low quality).

Independence of studies. When selecting studies for inclusion in the meta-analyses, we found that several articles contained potentially overlapping samples. For example, studies using data collected by the MyPersonality project potentially share parts of the same sample and data, and often investigate the same main characteristic. In general, a potential lack of independence exists between many of those studies, violating certain statistical assumptions of the meta-analysis. In efforts to resolve this issue, we followed recommendations from previous studies.[26,27] We considered studies as nonindependent if they met the following criteria: (1) each correlation was based on responses from overlapping sample subjects, (2) the main assessed characteristics were the same, (3) digital traces were extracted from the same social media platform, and (4) type of digital traces used to predict characteristics was the same or partly overlapping.

When studies were found to be nonindependent, the study with the most comprehensive set of digital traces was included in the analysis. In the case of nonindependent studies analyzing the same set of digital traces, the one with the larger sample size was included in the meta-analysis. In the case of studies including more than one effect size referring to the same psychological characteristic (e.g., Big 5 traits), we averaged the effect sizes to obtain a single effect size to ensure independence of the correlations entered into the meta-analysis.[28]

Strategy of analyses

For each study, an effect size was calculated. We used Pearson's r to express the relationship between digital traces and investigated outcomes. We chose not to transform correlations into Fisher's z scores for meta-analytic calculations because this transformation produces an upward bias in the estimation of mean correlation, which is usually higher than the downward bias produced by the use of untransformed correlations.[29]

When studies did not report Pearson's r, but instead reported alternative effect-size indicators (e.g., when characteristics were examined in dichotomous form by distinguishing individuals at low and high levels using validated or empirically derived cutoffs), reported effect sizes were converted to correlations. Area under the receiver operating characteristic curve statistics was first converted to Cohen's d,[30] and then converted from Cohen's d to r.[31] When studies provided specificity and sensitivity values, or positive predicted values and negative predicted values, or enough information was available for computing these statistics, we used this information to compute odds ratios,[32] then transformed odds ratios to Cohen's d,[33] and finally converted Cohen's d to correlations.[31]

When studies only reported the mean absolute error (MAE) and root mean square error (RMSE) statistics ($n = 7$), and thus did not provide enough information to compute

Table 1. Characteristics of Studies Included in the Systematic Review and Meta-Analyses

Study (effect)	*Main characteristic*	*Specific characteristic*	*Self-report*	r	N	*Source (quality)*	*Social media*	*Digital traces*
Bachrach et al.[18,a]	Personality	Big 5 traits	IPIP	0.40	5,000	Proceeding (low)	Facebook	Activity
Bai et al.[2] (1)	Psychological well-being	Life satisfaction	Urban and Rural Residents Social Attitudes Questionnaire	0.30	2,018	Repository (low)	Sina Weibo	Demographics, Activity, Language
Bai et al.[2] (2)	Social satisfaction	Income, social position, national economy, local economy, social justice, average satisfaction	(Same as above)	0.48	2,018	Repository (low)	Sina Weibo	Demographics, Activity, Language
Celli et al.[38] (1)	Personality	Big 5 traits	Big 5 Personality Test (BFI-10)	0.15	89	Proceeding (low)	Facebook	Pictures
Celli et al.[38] (2)	Personal values	Dominance and affect	Interpersonal Circumplex (IPIP-IPC-32)	0.16	89	Proceeding (low)	Facebook	Pictures
Chen et al.[53]	Personal values	Self-transcendence, self-enhancement, conservation, openness to change, hedonism	Portrait Value Questionnaire	0.39	799	Proceeding (high)	Reddit	Language
De Choudhury et al.[3]	Psychological well-being	Depression	CES-D and Beck Depression Inventory	0.48	476	Proceeding (low)	Twitter	Demographics, Activity, Language
Farnadi et al.[4] (1)[a]	Personality	Big 5 traits	IPIP	0.22	3,731	Journal (high)	Facebook	Demographics, Activity, Language
Farnadi et al.[4] (3)[a]	Personality	Big 5 traits	Big 5 Inventory—10	0.37	44	Journal (high)	Twitter	Demographics
Gao et al.[39]	Personality	Big 5 traits	44-Item Big 5 Personality Inventory	0.36	176	Proceeding (low)	Sina Weibo	Activity, Language
Garcia and Sikstrom[1]	Personality	Dark Triad, extraversion and neuroticism	Eysenck Personality Questionnaire Revised, Narcissistic Personality Inventory and Mach-IV	0.14	304	Journal (high)	Facebook	Language
Golbeck[54]	Copying style	Coping style	Ways of Coping Survey	0.59	105	Proceeding (low)	Twitter	Language
Golbeck et al.[7]	Personality	Big 5 traits	45-Item Big 5 Personality Inventory	0.57	167	Proceeding (low)	Facebook	Demographics, Activity, Language
Golbeck[40] (1)[a]	Personality	Big 5 traits	100-Item IPIP	0.35	127	Proceeding (low)	Facebook	Language
Golbeck[40] (2)[a]	Personality	Big 5 traits	100-Item IPIP	0.21	8,569	Proceeding (low)	Facebook	Language
Golbeck[40] (3)[a]	Personality	Big 5 traits	45-Item Big 5 Personality Inventory	0.24	69	Proceeding (low)	Facebook	Language
Gosling et al.[8]	Personality	Big 5 traits	TIPI	0.25	133	Journal (high)	Facebook	Activity
He et al.[52,a]	Self-monitoring	Self-monitoring skills	Snyder's Self-Monitoring Questionnaire	0.19	1,128	Journal (high)	Facebook	Language
Kern et al.[9,a]	Personality	Big 5 traits	IPIP	0.15	69,792	Journal (high)	Facebook	Language
Kleanthous et al.[41]	Personality	Big 5 traits	50-Item IPIP	0.15	62	Proceeding (low)	Facebook	Activity
Kosinski et al.[13] (1)[a]	Personality	Big 5 traits	IPIP	0.35	54,373	Journal (high)	Facebook	Likes
Kosinski et al.[13] (2)[a]	Psychological well-being	Satisfaction with life	Satisfaction with Life Scale	0.17	2,340	Journal (high)	Facebook	Likes
Kosinski et al.[13] (3)[a]	Intelligence	Intelligence	Raven's SPM	0.39	1,350	Journal (high)	Facebook	Likes
Kosinski et al.[13] (4)[a]	Substance use	Smokes cigarettes, drinks alcohol, and uses drugs	Online surveys—not specified	0.34	856–1,211	Journal (high)	Facebook	Likes
Kosinski et al.[19] (1)[a]	Personality	Big 5 traits	IPIP	0.17	9,515–45,565	Journal (high)	Facebook	Activity
Kosinski et al.[19] (2)[a]	Psychological well-being	Satisfaction with life	Satisfaction with life scale	0.33	311	Journal (high)	Facebook	Activity
Kosinski et al.[19] (3)[a]	Intelligence	Intelligence	Raven's SPM	0.2	395	Journal (high)	Facebook	Activity

(*continued*)

Table 1. (Continued)

Study (effect)	*Main characteristic*	*Specific characteristic*	*Self-report*	r	N	*Source (quality)*	*Social media*	*Digital traces*
Li et al.[14]	Personality	Big 5 traits	Chinese Version of the 44-Item Big 5 Personality Inventory	0.54	547	Journal (high)	Sina Weibo	Activity
Liu et al.[5,a]	Psychological well-being	Satisfaction with life	Satisfaction with life scale	0.15	1,124	Journal (high)	Facebook	Language
Liu et al.[25] (1)	Personality	Big 5 traits	IPIP	0.19	254	Proceeding (low)	Twitter	Language
Liu et al.[25] (2)	Personality	Big 5 traits	IPIP	0.12	429	Proceeding (low)	Twitter	Pictures
Markovikj et al.[20,a]	Personality	Big 5 traits	IPIP	0.67	250	Proceeding (low)	Facebook	Demographics, Activity, Language
Park et al.[10,a]	Personality	Big 5 traits	IPIP	0.38	4,824	Journal (high)	Facebook	Language
Preotiuc-Pietro et al.[42]	Personality	Dark triad	Dirty Dozen 12-Item Questionnaire	0.25	491	Proceeding (high)	Twitter	Language, Pictures
Qiu et al.[43]	Personality	Big 5 traits	44-Item Big 5 Personality Inventory	0.22	142	Journal (high)	Twitter	Language
Quercia et al.[21] (1)[a]	Personality	Big 5 traits	IPIP	0.08	2,165	Proceeding (high)	Facebook	Activity
Quercia et al.[21] (2)[a]	Self-monitoring	Self-monitoring skills	Snyder's Self-Monitoring Questionnaire	0.089	2,165	Proceeding (high)	Facebook	Activity
Schwartz et al.[6,a]	Personality	Big 5 traits	IPIP	0.35	18,177	Journal (high)	Facebook	Language
Schwartz et al.[24,a]	Psychological well-being	Depression	Average response to 7 Depression Facet Items from the Neuroticism Item Pool (IPIP)	0.39	1,000	Proceeding (low)	Facebook	Language
Schwartz et al.[49,a]	Psychological well-being	Satisfaction with life	Satisfaction with Life Scale and P.E.R.M.A Scale	0.30	440	Proceeding (low)	Facebook	Language
Settanni and Marengo[11]	Psychological well-being	Anxiety, depression, and stress	Adapted Version of DASS-21	0.32	201	Journal (high)	Facebook	Language
Skowron et al.[44]	Personality	Big 5 traits	44-Item Big 5 Personality Inventory	0.66	62	Proceeding (low)	Twitter and Instagram	Language, Pictures
Sumner et al.[45]	Personality	Big 5 traits and Dark triad	TIPI and Short Dark Triad (SD3) Questionnaire	0.2	616	Proceeding (low)	Twitter	Activity, Language
Thilakaratne et al.[46,a]	Personality	Big 5 traits	IPIP	0.38	1,000	Proceeding (low)	Facebook	Language
Tsugawa et al.[50] (3)	Psychological well-being	Depression	CES-D	0.32	209	Proceeding (high)	Twitter	Activity, Language
Wald et al.[47]	Personality	Big 5 traits	45-Item Big 5 Personality Index	0.68	537	Proceeding (low)	Facebook	Demographics, Activity, Language
Wang et al.[51]	Psychological well-being	Depression	Clinical psychological diagnosis	0.69	180	Proceeding (low)	Sina Weibo	Activity, Language
Wei and Stillwell[23,a]	Intelligence	Intelligence	Raven's SMP	0.27	7,000	Repository (low)	Facebook	Pictures
Wei et al.[48]	Personality	Big 5 traits	44-Item big personality inventory	0.40	949	Proceeding (low)	Sina Weibo	Activity, Language, Pictures
Youyou et al.[22,a]	Personality	Big 5 traits	100-Item IPIP	0.43	1,919	Journal (high)	Facebook	Likes

Note: Studies included in the meta-analyses are in bold.
[a]Study using MyPersonality data sets.
IPIP, International Personality Item Pool Questionnaire; SPM, Standard Progressive Matrices; TIPI, Ten Item Personality Inventory.

correlations, or results were not fully reported in the study ($n = 2$), we contacted the first author of the study to obtain any missing information. Missing information was obtained for one study ($n = 1$).

We conducted separate meta-analyses for each main characteristic (i.e., personality, psychological well-being, and intelligence). Meta-analyses were performed using a random-effects model as the true effect size was likely to vary in the individual studies; owing to the variety in data sources, study designs, and analytic approaches. Grubb's test was used to identify outliers. Heterogeneity of the studies' effect sizes included in each pooled analysis was evaluated by examination of (1) the chi-square Q statistic of heterogeneity, (2) the T^2 estimate of true between-study variance, and (3) the I^2 statistic of proportion of variation in observed effects due to the variation in true effects. Possible publication bias was evaluated by inspecting the funnel plot, by the statistical significance of the Begg and Mazumdar's adjusted rank correlation test[34] and Egger's test of the intercept,[35] Duval and Tweedie's trim-and-fill procedure,[36] and classic fail-safe *N*.

The effect of moderators on study effect sizes was measured by random-effects univariate meta-regressions using maximum-likelihood estimation. To obtain sufficiently robust coefficient estimates, we followed the suggestion by Fu et al.[37] and examined the effect of grouping variables only if at least four studies per group were available. We used a critical value of $\alpha = 0.05$ in our meta-regression analyses, but due to the low number of studies, effects approaching statistical significance ($p < 0.10$) are commented as suggestions of possible links worthy of being explored by future researches.

Results

Overview of studies

We found 38 articles, resulting in 50 different effect sizes (Table 1). Information about all selected studies is shown in Tables 1 and 2.

Overall, we found three characteristics for which at least three studies were published, namely personality (26 articles,[1,4,6–10,13,14,18–22,25,38–48] including 30 effect sizes), psychological well-being (10 articles,[2,3,5,11,13,19,24,49–51] including 10 effect sizes), and intelligence (3 articles,[13,19,23] including 3 effect sizes). Other characteristics for which we found fewer than three studies were social satisfaction,[2] substance use,[13] self-monitoring skills,[21,52] personal values,[38,53] and coping style.[54]

Meta-analyses were performed on characteristics that were reported in at least three studies. After inspection of studies for nonindependence, we selected a subset of 25 articles, including 30 independent effect sizes, about the three main characteristics, namely personality ($n = 18$), psychological well-being ($n = 9$), and intelligence ($n = 3$) (Table 1). Grubb's test failed to identify any outliers, resulting in no further studies being excluded. Results of meta-analyses are reported below.

TABLE 2. INDEPENDENT CHARACTERISTICS OF ARTICLES INCLUDED IN THE META-ANALYSIS OF DIGITAL TRACES, SOCIAL MEDIA, AND PSYCHOSOCIAL CHARACTERISTICS: STUDY CHARACTERISTICS

Study characteristics	*Frequency*
Effect sizes	$n = 50$
Personality	30
Big 5 traits	29
Dark-Triad	3
Psychological well-being	10
Depression	4
Emotional distress	1
Satisfaction with life	5
Intelligence	3
Other psychosocial characteristics	6
Social satisfaction	1
Personal values	2
Coping style	1
Self-monitoring skills	2
Substance use	1
Social media platform	$n = 50$
Facebook	33
Twitter	10
Sina-Weibo	6
Instagram	1
Reddit	1
Type of digital traces	$n = 50$
Language	31
Activity statistics	21
Likes	7
Demographics	8
Pictures	5
Analytic approach	$n = 50$
Multiple features	14
Single features	36
MyPersonality studies	26
Sample size	$n = 49$
Less than 200	10
201–500	12
501–1,000	6
1,001–5,000	12
5,001–10,000	3
More than 10,000	6
Publication source	$n = 38$
Online repository	2
Proceedings	16
Journal	20

Meta-analyses

Personality

Mean effect size. To establish the magnitude of the association between digital traces and personality, we analyzed 18 independent effect sizes. The estimated meta-analytic correlation was 0.34, 95% CI [0.27–0.34] (Fig. 2), and this effect was significantly greater than 0, $z = 9.58$, $p < 0.001$. Q test for heterogeneity was significant: Q (17) = 318.33 ($p < 0.001$). There was low true heterogeneity between studies, $T^2 = 0.02$ ($T = 0.14$), and the observed dispersion of effect sizes was mostly due to true heterogeneity ($I^2 = 94.66$).

Publication bias. First, we inspected the funnel plot (Fig. 3), plotting the included studies' effect size against its standard error. The funnel plot was symmetrical, suggesting lack of publication bias. Trim-and-fill analysis suggested that no studies were missing on the left side of the mean effect. The *p*-values of Begg and Mazumdar's test and Egger's test were $p = 0.52$ and $p = 0.43$, indicating no significant evidence of publication bias. The result of classic fail-safe *N* suggested

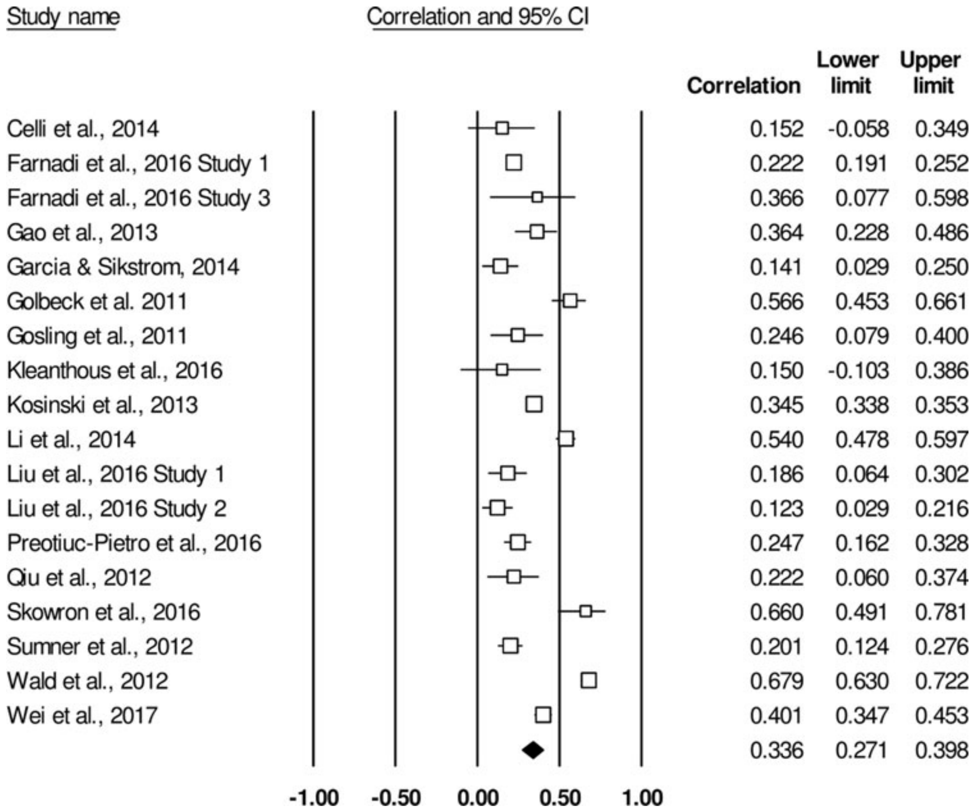

FIG. 2. Forest plot of personality study-average effect sizes by weight.

that 9,638 null reports would be required in order for the combined two-tailed p-value to exceed the alpha level of 0.05. The fail-safe N value was larger than 100, corresponding to the recommended rule-of-thumb limit of 5k+10.[55] The results of these four tests indicated that it is unlikely that publication bias poses a significant threat to the validity of the findings reported in the current analysis.

Moderator analyses. We examined the following moderating effects. (1) Privacy versus public oriented social media platform, (2) multiple versus single types of digital traces, (3) use of user demographics, (4) use of activity statistics, (5) use of language-based features, (6) use of pictures, and (7) study quality. Results of univariate meta-regressions, shown in Table 3, indicated an increase in strength of association between digital traces and personality when studies examined multiple types of digital traces compared with only one type ($K=18$, $\beta=0.19$, $p<0.05$). Use of demographic statistics for prediction purposes was also associated with an increase in correlation strength between digital traces and personality ($K=18$, $\beta=0.23$, $p<0.05$). The remaining moderators did not show significant effects.

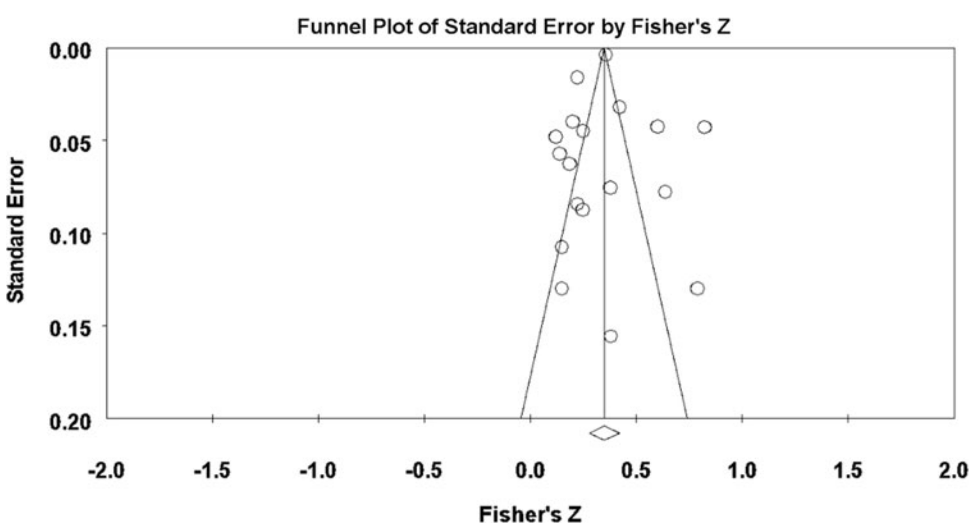

FIG. 3. Funnel plot displaying effect sizes for personality by SEs. SE, standard error.

Table 3. Results of Univariate Meta-Regressions: Factors Moderating Effect Size for Personality (K = 18) and Psychological Well-Being (K = 9)

	T^2	R^2	β	*SE*	*95% CI*	p
Personality						
Study quality (high vs. low)	0.04	0.04	–0.08	0.10	–0.28 to 0.11	0.41
Private versus public settings	0.04	0.00	0.01	0.1	–0.19 to 0.21	0.95
Multiple versus single types of digital traces	0.03	0.19	0.19	0.09	0.01 to 0.38	0.04
User demographics (yes vs. no)	0.03	0.22	0.23	0.11	0.01 to 0.44	0.04
User activity statistics (yes vs. no)	0.03	0.16	0.15	0.09	–0.04 to 0.33	0.12
Language (yes vs. no)	0.04	0.03	0.1	0.11	–0.11 to 0.30	0.37
Pictures (yes vs. no)	0.04	0.01	–0.02	0.11	–0.24 to 0.20	0.83
Psychological well-being						
Study quality (high vs. low)	0.04	0.04	–0.08	0.10	–0.28 to 0.11	0.41
Private versus public settings	0.02	0.31	–0.18	0.10	–0.37 to 0.01	0.06
Multiple versus single types of digital traces	0.02	0.31	0.18	0.10	0.01 to 0.38	0.06
User activity statistics (yes vs. no)	0.03	0.16	0.15	0.09	–0.04 to 0.33	0.12

Psychological well-being

Mean effect size. The magnitude of the association between digital traces and psychological well-being was analyzed by summarizing nine independent effect sizes. The estimated meta-analytic correlation was 0.37, 95% CI [0.28–0.45] (Fig. 4), and this effect was significantly greater than 0, $z = 7.54$, $p < 0.001$. Q test for heterogeneity was significant: Q $(8) = 124.67$ ($p < 0.001$). There was relatively low true heterogeneity between studies, $T^2 = 0.02$ ($T = 0.14$), and the observed dispersion of effect sizes was mostly due to true heterogeneity ($I^2 = 93.58$).

Publication bias. Inspection of funnel plot (Fig. 5) and trim-and-fill analysis suggested that no studies were missing on the left side of the mean effect. The *p*-values of Begg and Mazumdar's test and Egger's test were $p = 0.37$ and $p = 0.07$, indicating low probability of publication bias. The result of classic fail-safe *N* suggested that 1,618 null reports would be required in order for the combined two-tailed *p*-value to exceed the alpha level of 0.05. The fail-safe *N* value was larger than the recommended rule-of-thumb limit of 55. Overall, results did not suggest existence of significant publication bias.

Moderator analyses. We examined the following moderating effects. (1) Multiple versus single sources of digital traces, (2) type of social media platform (private vs. public), (3) use of activity statistics, and (4) study quality. Remaining categorical moderators were not tested because they did not reach the per-group minimum value of four distinct studies. Results of univariate meta-regressions (Table 3) seem to indicate a relevant increase in the effect size of the association between digital traces and psychological well-being when using multiple types of digital traces compared with use of only one type (K = 9, $\beta = 0.18$, $p < 0.10$). In addition, when comparing studies conducted on private social media platform (e.g., Facebook) with those conducted on public platforms (e.g., Twitter), a relevant difference in effect size in favor of public platforms emerged ($\beta = –0.18$, $p < 0.10$), even if the effect did not reach proper significance.

However, given the perfect collinearity between the variables concerning private/public platforms and multiple/single types of digital traces, a univocal interpretation of these moderator effects is not possible. A larger and more differentiated sample of studies will permit to ascertain both the presence of a significant impact of the type of platform

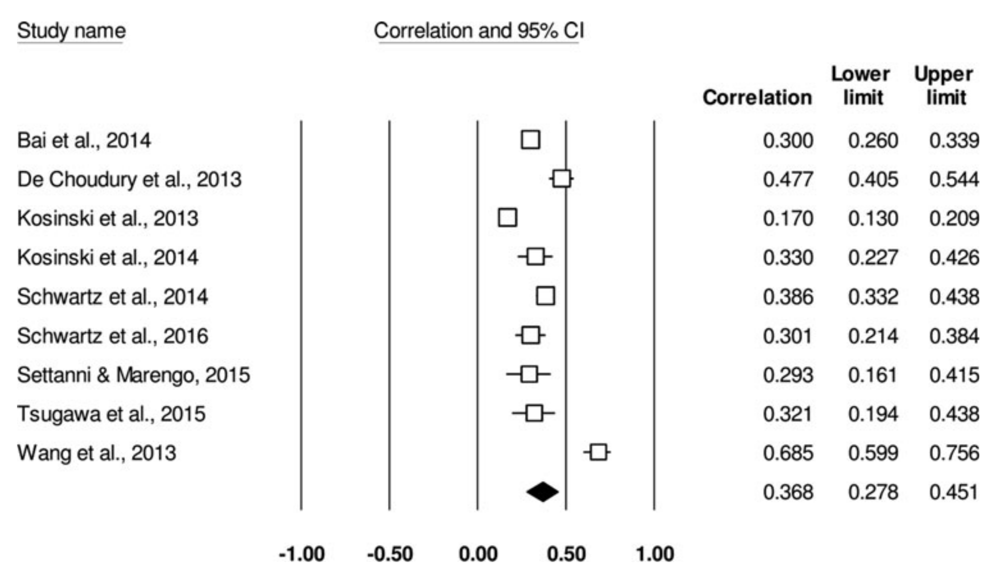

FIG. 4. Forest plot of psychological well-being study-average effect sizes by weight.

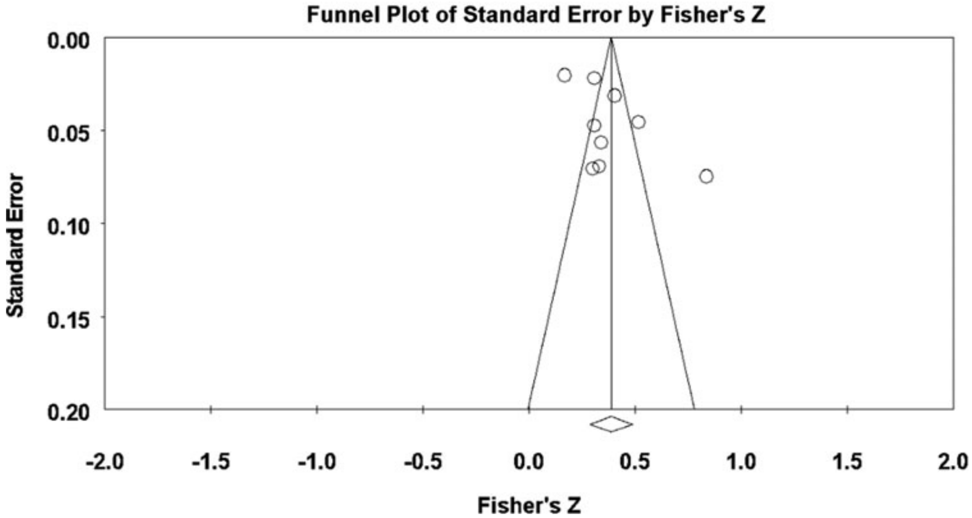

FIG. 5. Funnel plot displaying effect sizes for psychological well-being by SEs.

and distinguish between the effects of multiple versus single types of digital traces and type of platform analyzed. Remaining moderators did not show relevant effect.

Intelligence

Mean effect size. The magnitude of the association between digital traces and intelligence was analyzed by summarizing effects presented in three studies. The estimated meta-analytic correlation was 0.29, 95% CI [0.19–0.38] (Fig. 6), and this effect was significantly greater than zero, $z = 5.65$, $p < 0.001$. Q test for heterogeneity was significant: Q (2) = 24.01 ($p < 0.001$). There was low between-study heterogeneity, $T^2 = 0.01$ ($T = 0.09$), and the observed dispersion of effect-sizes was mostly due to true heterogeneity ($I^2 = 91.67$).

Publication bias. On examination, funnel plot (Fig. 7) was found to be symmetrical, suggesting no publication bias. In addition, trim-and-fill analysis suggested that no studies were missing on the left side of the mean effect. The Begg and Mazumdar's ($p = 0.99$) and Egger's test ($p = 0.84$) and the result of classic fail-safe N suggested (463 null reports required to exceed the alpha level of 0.05) lack of publication bias.

Discussion

To our knowledge, this is the first meta-analysis summarizing results from studies investigating the use of digital traces collected from social media to predict psychological and behavioral characteristics. Our main aim was to determine the mean effect size of associations between digital traces from social media and specific individual characteristics. We found 38 articles using features automatically extracted from digital traces of human behavior on social media to predict different psychosocial characteristics. Meta-analyses were conducted on characteristics investigated by at least three independent studies; personality, psychological well-being, and intelligence. Overall, we found the majority of reported associations between features extracted from digital traces and investigated characteristics to be at least of moderate strength. Significant associations with digital traces were found for each of the most investigated characteristics, with mean correlation values (Pearson's r) ranging from 0.29 (intelligence) to 0.37 (psychological well-being).

Included effect sizes showed low-to-moderate dispersion that was mostly due to true differences across studies. Given the presence of heterogeneity of effects among studies, potential moderator effects were analyzed with respect to the following possible sources of variation: type of social media platform (private vs. public), type of extracted features, and analytical approaches (e.g., use of multiple vs. single types of digital traces). Our hypothesis was that each of these factors and their interactions could contribute to the overall heterogeneity of effects. Unfortunately, given the small number of studies included in the meta-analyses, we were able to

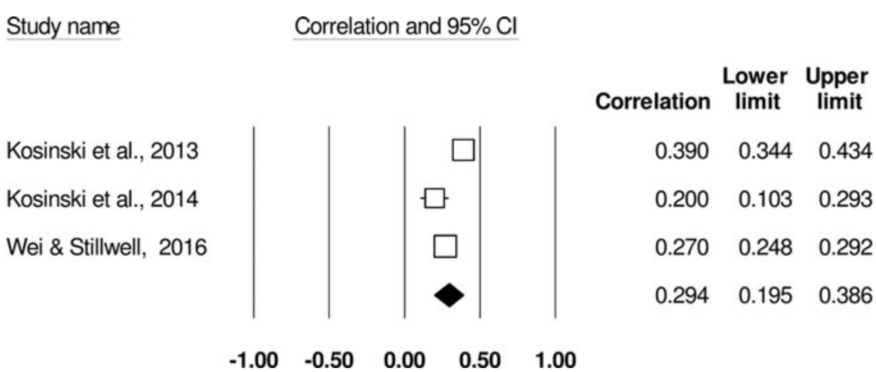

FIG. 6. Forest plot of intelligence study-average effect sizes by weight.

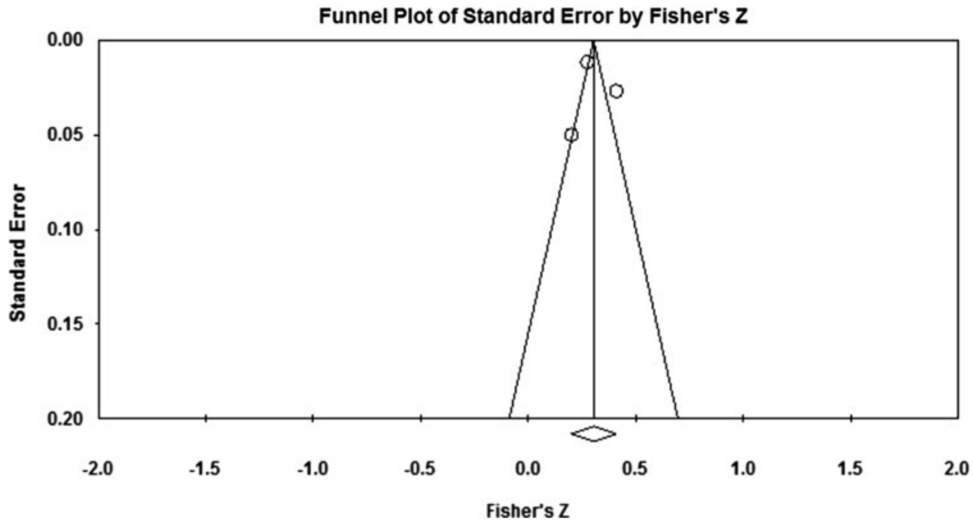

FIG. 7. Funnel plot displaying effect sizes for intelligence by SEs.

perform moderator analyses for only two characteristics; personality and psychological well-being. Moreover, we were able to investigate the influence of only a subset of the moderators, and it was not possible to test the influence of interaction effects between moderators.

Our results indicate that the association between digital traces, and both personality and well-being, was stronger when multiple types of digital traces were analyzed. Regarding the type of extracted features, use of demographics extracted from social media positively affected the strength of the relationship between personality and digital traces, suggesting the opportunity to include them in models aimed at increasing the predictive power of digital traces. Furthermore, the type of social media platform (public vs. private) did not affect the strength of association with personality, while digital traces extracted from private platforms were less strongly associated with psychological well-being.

Overall, analysis of moderators outlined that a significant part of the effect size heterogeneity can be traced back to the amount of digital traces included in the studies: generally, higher effect sizes have been achieved by studies, including multiple types of digital traces. We hypothesize that future studies will confirm this relationship; hence, to reach a higher predictive power, scholars should collect data from a large set of different digital traces, possibly combining different types of data (e.g., pictures and text) from different social media platforms.

As noted in the Introduction, most of the reviewed studies focused on predicting individual characteristics without providing explanations or hypotheses regarding the existing relationships between specific digital traces and outcomes. This approach is quite common in computer science, while relatively novel among other disciplines. As this approach becomes more common in psychology and social sciences, we expect that findings of predictive studies may significantly contribute to the refinement of existing, and the building of new theories.

Conclusions

The present meta-analysis demonstrates that digital traces extracted from social media can be used to infer specific psychological characteristics. The presence of significant associations between digital traces and psychosocial characteristics and the lack of relevant differences in the strength of these associations indicate that records of digitally mediated behaviors from social media can be used to study and predict theoretically distant psychosocial characteristics with comparable accuracy. The relationship between digital traces of online behavior and psychological characteristics is quite strong, apparently stronger than the association found by scholars studying the link between personality and offline behaviors.[56,57] We expect the accuracy to grow in the future due to the ongoing transition from the use of traditional analytic approaches toward a more pervasive use of data mining techniques,[58] and the emergence of new techniques to extract meaningful information from visual data,[59] which is especially important, given the current shift in content sharing on social media, from text to photos and videos.[60] These methodological improvements will hopefully help this research area become more mainstream among social scientists, which in turn will favor the theoretical reflection regarding the relationship between actual online behaviors and individual characteristics.

Results from the present study have implications on the development of tools allowing for the unobtrusive assessment of psychological characteristics of social media users, which in turn can be beneficial for a variety of purposes, including commercial applications (e.g., user-tailored advertising and online experiences) and health-related purposes (e.g., early detection of individuals at risk for depression, longitudinal tracking of mental well-being trends).

However, possible questionable uses of these tools also exist: recently, newspapers reported cases showing the feasibility and the efficacy of targeting political messages on the basis of unintentionally disclosed information on social media[61,62] or targeting ads on the basis of users' emotional state.[63] The risks associated with the application of these new techniques to specific areas and subjects should be carefully considered by scholars.

Author Disclosure Statement

No competing financial interests exist.

References

1. Garcia D, Sikström S. The dark side of Facebook: semantic representations of status updates predict the Dark Triad of personality. Personality and Individual Differences 2014; 67:92–96.
2. Bai S, Gao R, Hao B, et al. Identifying social satisfaction from social media. arXiv Preprint 2014; arXiv:1407.3552.
3. De Choudhury M, Gamon M, Counts S, et al. (2013) Predicting depression via social media. In *Proceedings of 7th International AAAI Conference on Weblogs Social Media*. Boston, MA, July 8–10.
4. Farnadi G, Sitaraman G, Sushmita S, et al. Computational personality recognition in social media. User Modeling and User-Adapted Interaction 2016; 26:109–142.
5. Liu P, Tov W, Kosinski M, et al. Do Facebook status updates reflect subjective well-being? Cyberpsychology, Behavior, and Social Networking 2015; 18:373–379.
6. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 2013; 8:e73791.
7. Golbeck J, Robles C, Turner K. (2011) Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. Vancouver, Canada: ACM, pp. 253–262.
8. Gosling SD, Augustine AA, Vazire S, et al. Manifestations of personality in online social networks: self-reported Facebook-related behaviors and observable profile information. Cyberpsychology, Behavior, and Social Networking 2011; 14:483–488.
9. Kern ML, Eichstaedt JC, Schwartz HA, et al. The online social self an open vocabulary approach to personality. Assessment 2013; 21:158–169.
10. Park G, Schwartz HA, Eichstaedt JC, et al. Automatic personality assessment through social media language. Journal of Personality and Social Psychology 2015; 108:934–952.
11. Settanni M, Marengo D. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. Frontiers in Psychology 2015; 6:1045.
12. Farrell D, Petersen JC. The growth of internet research methods and the reluctant sociologist. Sociological Inquiry 2010; 80:114–125.
13. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 2013; 110:5802–5805.
14. Li L, Li A, Hao B, et al. Predicting active users' personality based on micro-blogging behaviors. PLoS One 2014; 9:e84997.
15. Madden M, Fox S, Smith A, et al. (2007) Digital footprints. Pew Research Center. www.pewinternet.org/2007/12/16/digital-footprints (accessed November 1, 2017).
16. Andrew SH, Ungar LH. Data-driven content analysis of social media: a systematic overview of automated methods. The ANNALS of the American Academy of Political and Social Science 2015; 659:78–94.
17. Galit S. To explain or to predict? Statistical Science 2010; 25:289–310.
18. Bachrach Y, Kosinski M, Graepel T, et al. (2012) Personality and patterns of Facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference 2012*. Evanstone, IL: ACM, pp. 24–32.
19. Kosinski M, Bachrach Y, Kohli P, et al. Manifestations of user personality in website choice and behaviour on online social networks. Machine Learning 2014; 95:357–380.
20. Markovikj D, Gievska S, Kosinski M, et al. (2013) Mining facebook data for predictive personality modeling. In *Proceedings of the 7th international AAAI conference on Weblogs and Social Media*. Boston, MA: ICWSM, pp. 23–26.
21. Quercia D, Lambiotte R, Stillwell D, et al. (2012) The personality of popular facebook users. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. Seattle, WA: ACM, pp. 955–964.
22. Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences 2015; 112:1036–1040.
23. Wei X, Stillwell D. How smart does your profile image look? Intelligence estimation from social network profile images. arXiv Preprint 2016; arXiv:1606.09264.
24. Schwartz HA, Eichstaedt J, Kern ML, et al. (2014). Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pp. 118–125.
25. Liu L, Preotiuc-Pietro D, Samani ZR, et al. (2016). Analyzing personality through social media profile picture choice. In *Proceedings of the 10th International AAAI Conference Web & Social Media, 2011*. Cologne, Germany: ICWSM. pp. 211–220.
26. Hunter J, Schmidt F, Jackson G. (1982) *Meta-analysis: cumulating research findings across studies.* Beverly Hills, CA: Sage.
27. Sheppard BH, Hartwick J, Warshaw PR. The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research. Journal of Consumer Research 1988; 15:325–343.
28. Hunter JE, Schmidt FL. Dichotomization of continuous variables: the implications for meta-analysis. Journal of Applied Psychology 1990; 75:334–349.
29. Schmidt FL, Hunter JE. (2014). *Methods of meta-analysis: correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.
30. Ruscio J. A probability-based measure of effect size: robustness to base rates and other factors. Psychological Methods 2008; 13:19–30.
31. Rosenthal R. (1994). Parametric measures of effect size. In Cooper H, Hedges LV, eds. *The handbook of research synthesis*. New York, NY: Sage, pp. 239.
32. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology 2003; 56:1129–1135.
33. Borenstein M, Hedges LV, Higgins J, et al. Meta-regression. Introduction to Meta-analysis 2009:187–203.
34. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994:1088–1101.
35. Sterne JA, Matthias E. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. Journal of Clinical Epidemiology 2001; 54:1046–1055.
36. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000; 56:455–463.
37. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews 2014:254.
38. Celli F, Bruni E, Lepri B. (2014) Automatic personality and interaction style recognition from facebook profile pictures.

In *Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando, FL: ACM.
39. Gao R, Hao B, Bai S, et al. (2013) Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM Conference on Recommender Systems*. Hong Kong: ACM.
40. Golbeck J. Predicting personality from social media text. AIS Transactions on Replication Research 2016; 2:2.
41. Kleanthous S, Herodotou C, Samaras G, et al. Detecting personality traces in users' social activity. In *International Conference on Social Computing and Social Media*. Toronto, Canada: Springer International Publishing, 2016.
42. Preotiuc-Pietro D, Carpenter J, Giorgi S, et al. (2016) Studying the Dark Triad of personality through Twitter behavior. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM.
43. Qiu L, Lin H, Ramsay J, et al. You are what you tweet: personality expression and perception on Twitter. Journal of Research in Personality 2012; 46:710–718.
44. Skowron M, Tkalcic M, Ferwerda B, et al. (2016) Fusing social media cues: personality prediction from twitter and instagram. In *Proceedings of the 25th international conference companion on world wide web*. International World Wide Web Conferences Steering Committee, pp. 107–108.
45. Sumner C, Byers A, Boochever R, et al. (2012) Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *Machine Learning and Applications ICMLA, 2012 11th International Conference on. Vol. 2.* Boca Raton, FL, IEEE, pp. 386–393.
46. Thilakaratne M, Weerasinghe R, Perera S. (2016) Knowledge-driven approach to predict personality traits by leveraging social media data. In *Web Intelligence WI, 2016 IEEE/WIC/ACM International Conference*. Omaha, NE: IEEE, pp. 288–295.
47. Wald R, Khoshgoftaar T, Sumner C. (2012) Machine prediction of personality from Facebook profiles. In *Information Reuse and Integration IRI, 2012 IEEE 13th International Conference*. Las Vegas, NV: IEEE, pp. 109–115.
48. Wei H, Zhang F, Yuan NJ, et al. (2017) Beyond the words: predicting user personality from heterogeneous information. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. Cambridge, UK: ACM, pp. 305–314.
49. Schwartz HA, Sap M, Kern ML, et al. Predicting individual well-being through the language of social media. Biocomputing 2016: Proceedings of the Pacific Symposium 2016; 21:516–21.
50. Tsugawa S, Kikuchi Y, Kishino F, et al. (2015) Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, South Korea: ACM, pp. 3187–3196.
51. Wang X, Zhang C, Ji Y, et al. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg 2013:201–213.
52. He Q, Glas CA, Kosinski M, et al. Predicting self-monitoring skills using textual posts on Facebook. Computers in Human Behavior 2014; 33:69–78.
53. Chen J, Hsieh G, Mahmud JU, et al. (2014) Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, pp. 405–414.
54. Golbeck J. (2016) Detecting coping style from Twitter. In *International Conference on Social Informatics*. Springer International Publishing, pp. 454–467.
55. Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin 1979; 86:638.
56. Meyer GJ, et al. Psychological testing and psychological assessment: a review of evidence and issues. American Psychologist 2001; 56:128.
57. Roberts BW, Kuncel NR, Shiner R, et al. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2007; 2:313–345.
58. Kosinski M, Wang Y, Lakkaraiu H, et al. Mining big data to extract patterns and predict real-life outcomes. Psychological Methods 2016; 21:493.
59. Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: a review. Neurocomputing 2016; 187:27–48.
60. Statista. The Most popular mobile social networking apps in the United States. 2017. www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/ (accessed July 2, 2017).
61. Confessore N, Hakim D. Data firm says 'secret sauce' aided Trump: many Scoff. New York Times, March 6, 2017. www.nytimes.com/2017/03/06/us/politics/cambridge- analytica.html?_r=0 (accessed June 25, 2017).
62. Cadwalladr C. The great British Brexit robbery: how our democracy was hijacked. The Guardian, May 7, 2017. www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (accessed July 2, 2017).
63. Levin S. Facebook told advertisers it can identify teens feeling 'insecure' and 'worthless'. The Guardian, May 1, 2017. www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens (accessed June 26, 2017).

Address correspondence to:
Danny Azucar
Department of Psychology
University of Turin
10124 Via Verdi 10
Turin 10124
Italy

E-mail: dazucarc@unito.it