8-25-2021

# The Use of Machine Learning in Assessing Suicide Risk: A Meta-analysis

Rachael Kang
rmk9710@rit.edu

Department of Psychology, College of Liberal Arts

Rochester Institute of Technology

# The Use of Machine Learning in Assessing Suicide Risk: A Meta-analysis

by

Rachael Kang

A Thesis in

Experimental Psychology

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

August 25, 2021

We approve the thesis of Rachael Kang:

| | |
|---|---|
| Esa M. Rantanen, Ph.D.<br>Associate Professor, Rochester Institute of Technology<br>Faculty Adviser and Chair of the Thesis Committee | Date |
| Lindsay S. Schenkel, Ph.D.<br>Associate Professor, Rochester Institute of Technology<br>Reader | Date |
| Clark G. Hochgraf, Ph.D.<br>Associate Professor, Rochester Institute of Technology<br>Reader | Date |
| Christopher M. Homan, Ph.D.<br>Associate Professor, Rochester Institute of Technology<br>Reader | Date |
| Eric A. Youngstrom, Ph.D.<br>Professor, University of North Carolina at Chapel Hill<br>Reader | Date |

**Acknowledgements**

I would like to acknowledge the extensive support and help from my wonderful thesis advisor, Dr. Esa Rantanen, as well as my esteemed committees members Drs. Lindsay Schenkel, Christopher Homan, Clark Hochgraf, and Eric Youngstrom. It was under the guidance and mentorship of my committee members and advisor that I was able to successfully undertake and complete this thesis. I would also like to thank my research assistants Erin Devilbiss and Wilson Jacobs, two incredibly bright and enthusiastic students who belong to the UNC-Chapel Hill chapter of the Helping Give Away Psychological Science (HGAPS) organization.

Finally, I would like to thank my friends Sylvia and Katie for their moral support throughout this endeavor as well as my mom and dad who pushed me to be the very best, like no one ever was.

Financial support for this research was provided by the College of Liberal Arts Research Fund grant at the Rochester Institute of Technology, Department of Psychology.

**Abstract**

Suicide is a devastating act in which a person takes their own life. Decades of research into suicide have identified a myriad of risk factors that have been used to create assessments of suicide risk and suicidality. However, more recent research has suggested that these identified risk factors may have no better predictive ability than chance, perhaps because suicide is actually a multi-dimensional, multi-faceted construct that has been viewed too simplistically for prediction's sake. To try and better appreciate the complex nature of suicide while also increasing prediction accuracy, researchers have turned to machine learning. This study sought to meta-analyze the predictive ability of machine learning in predicting suicide risk. A multi-level, mixed effects meta-analytic model returned a significant model with an effect size of $g = 1.36$ ($p < 0.0001$), but with a significant amount of heterogeneity ($Q(285\ df) = 66361.51$, $p < 0.0001$). A fully augmented model using three moderators (algorithm type, data source type, and suicide definition) accounted for a significant portion of the variance and also returned a statistically significant model. Meta-regression models showed that algorithm type had a statistically significant effect on the reported effect sizes while data source type and suicide definition did not return significant models. The results of this analysis found not only that machine learning indeed has a significant impact on the accuracy of predicting suicide, but also that the type of algorithm used has a significant impact on the reported accuracies as well. However, high within and between study hetereogeneity warrants more research into other potential moderating variables.

## Contents

## List of Figures

## List of Tables

# The Use of Machine Learning in Assessing Suicide Risk: A Meta-analysis

## Suicide

Suicide is an act in which a person is successfully able to take their own life. The World Health Organization reported that suicide accounted for nearly 1.5% of all deaths worldwide, making it the world's 20th leading cause of death, the second leading cause of death in 15 to 29 years olds, and the number one leading cause of death in children under 15 (Kassebaum et al., 2017; Apter, Bursztein, Bertolote, Fleischmann, & Wasserman, 2009; WHO, 2017).

In the past two decades, cases of suicide have increased around the world in both adults and children. In a longitudinal meta-analysis of suicide studies published between 1984 and 2014, high-risk suicide cases (i.e., cases in which there is a high probability that a person will attempt suicide) in adults increased from 2.0% of cases in 1984 to 51.4% of cases in 2014 (J. D. Ribeiro, Huang, Fox, & Franklin, 2018). Of these cases, 47.7% of patients attempted suicide and 33.6% of patients died by suicide. Further, two studies reported that, in the United States (US) alone, 44 states saw an increased rate of suicide with 25 states experiencing an over 30% increase (Rossen, Hedegaard, Khan, & Warner, 2018; Stone et al., 2018). And with these findings of increasing suicide rates there have been, unfortunately, other studies that have found that there have been no statistically significant decreases in suicide rates in youths around the world aged 10 to 19 years old over the past two decades (Kassebaum et al., 2017; Kolves & De Leo, 2014; Kõlves & De Leo, 2016).

Certain demographic groups appear to be more affected by suicide than other groups. Socio-economic status (SES) as well as economic recessions and growths have affected suicide rates. In the US, during periods of recession (e.g., Great Depression, New Deal, Oil Crisis, Double Dip), there was a sharp increase in the rates of suicide in both men and women across all age groups as the socioeconomic statuses of the population decreased (Luo, Florence, Quispe-Agnoli, Ouyang, & Crosby, 2011). Further, across the world, those who lived in rural, under-developed, or economically depressed areas have had higher rates of suicide

than those in areas that were developed and thriving (Cheung, 2014; X. Huang, Ribeiro, Musacchio, & Franklin, 2017; Kolves & De Leo, 2014; WHO, 2014).

Though suicide affects all sexes and age groups, patterns within sexes and ages have emerged (X. Huang et al., 2017). Sex was shown to be able to significantly predict suicide above other demographic variables (X. Huang et al., 2017). Males had higher rates of suicide than females across all age groups (Kolves & De Leo, 2014; Kõlves & De Leo, 2016; NIMH, 2019; WHO, 2014). Additionally, in places where there was a sudden economic depression, the rates of suicide in males increased significantly more than the rates of suicide in females (Kolves & De Leo, 2014; Luo et al., 2011). Furthermore, in different age groups there have also been varying rates of suicide. Children under 10 years old had the lowest rates of suicide, young adults between the ages of 15 and 20 had the next highest rates, and the highest rates of suicide were found in those who were 65 and older (Cheung, 2014; Soole, Kõlves, & De Leo, 2014; WHO, 2014). However, the highest rates of suicidal thoughts occurred in those who were 18 to 25 years old, and suicidal thoughts have been identified as a heightened risk factor for suicide (Millner & Nock, 2020; NIMH, 2019).

Research has also revealed that suicide rates vary between different racial groups (X. Huang et al., 2017). In the US and Australia, the indigenous populations have had the highest rates of suicide compared to the non-indigenous population (Cheung, 2014; NIMH, 2019). Specifically in the US, those who identified as white have had higher rates of suicide than those of African-American, Asian, and Hispanic descent (Bolton & Robinson, 2010; Cheung, 2014; NIMH, 2019).

Research suggests that there are definite patterns of suicide rates found within different demographic groups. SES and sex seem to be strong indicators of suicide followed by age and race. Researchers believed that these demographic patterns could be used to assess suicide risk in an individual or a population of people (X. Huang et al., 2017). Thus, these findings beget the question as to whether or not demographic information can be considered a suicide risk factor.

## Suicide Risk Factors

A risk factor is defined as a measurable characteristic in an individual or population that can be used not only to predict a future outcome, but also to split the population into two groups such as a low risk group and a high risk group (Kraemer et al., 1997). Though the conventional belief is that suicide is the result of an underlying mental illness, the risk factors associated with suicide encompass a myriad of sociocultural and psychological constructs. As a result, suicide risk factors can be complex and difficult to assess.

Shneidman (1993) suggested that suicide was the result of "psychache" (i.e., psychological pain). Research supported this sentiment as mental illnesses have been reported to be a significant risk factor for suicidal behavior and action (CDC, 2019; NIMH, 2019; J. D. Ribeiro et al., 2018; Stone et al., 2018; WHO, 2014). However, it would be erroneous to assume all mental illnesses carry an equal risk of suicide. For example, researchers have reported that nearly 50% of patients with schizophrenia attempted suicide while lifetime suicide rates for major depressive disorder was about 3.4% (Blair-West, Cantor, Mellsop, & Eyeson-Annan, 1999; Meltzer, 2002). Additionally, research has also demonstrated a strong, positive correlation between non-suicidal self-injury (i.e., self-injury with no intent to die), suicidal self-injury (i.e., self-injury with some intent to die), suicidal ideation, suicide plan, and a history of suicide attempts to suicide related deaths (Millner & Nock, 2020; CDC, 2019; NIMH, 2019; J. Ribeiro et al., 2016; Scott, Pilkonis, Hipwell, Keenan, & Stepp, 2015; Soole et al., 2014; Wenzel et al., 2011; WHO, 2014; Wilkinson, Kelvin, Roberts, Dubicka, & Goodyer, 2011). While research demonstrates that mental illnesses is highly associated with increased suicide risk, there are other risk factors that contribute to increased risk of suicide.

Outside of mental illnesses, sociocultural factors can also impact suicide risk. Substance abuse and alcohol dependence have been shown to be a risk factor for suicide (CDC, 2019; Stone et al., 2018; NIMH, 2019; WHO, 2014). Additionally, some kind of major, precipitating life event such as loss, injury, divorce, or economic depression have also been linked to increased risk of suicide (Bolton & Robinson, 2010; Soole et al., 2014). In the US, the Na-

tional Institute of Mental Health (NMIH) and the Center for Disease Control (CDC) have identified several other potential suicide risk factors: impulsive or aggressive tendencies, cultural/religious beliefs of the nobility of suicide, access to lethal methods of suicide, availability of mental health resources, history of child maltreatment and abuse, and exposure to other suicides (NIMH, 2019; CDC, 2019).

Despite extensive research into suicide risk factors, however, recent research has demonstrated that the predictive ability of the identified risk factors listed above might be no better than chance (Franklin et al., 2017; Linthicum, Schafer, & Ribeiro, 2019; J. Ribeiro et al., 2016; J. D. Ribeiro et al., 2018). Linthicum et al. (2019) and Franklin et al. (2017) suggested that no one risk factor or risk factor category stood out as being substantially stronger than the other at predicting suicide. Further, there has also been contradicting evidence on the effectiveness of certain risk factors in predicting suicide. For example, even though research has purported that mental illness is a significant suicide risk factor, several studies have shown that approximately 50% of suicide victims had no known mental illness diagnoses at the time of death (Bolton & Robinson, 2010; Soole et al., 2014; Stone et al., 2018). X. Huang et al. (2017) found no relationship among divorce, marriage, or religion in suicide, and Soole et al. (2014) found that only approximately one-third of children who committed suicide experienced any type of abuse. These contradictions just begin to demonstrate the complexity of assessing suicide risk and suicide risk factors.

## Current Methods of Suicide Assessment

Because previous research has suggested that suicide rates are higher in populations with a history of mental illnesses, previous suicide attempts, and other health complications (e.g., chronic pain or chronic health problems), current methods of suicide assessment give greater weight to these risk factors (J. D. Ribeiro et al., 2018; Scott et al., 2015; Stone et al., 2018; Wenzel et al., 2011; WHO, 2014; Wilkinson et al., 2011). Historically, there have been three generations of suicide assessment: (1) unassisted clinician prediction, (2) standardized scales, and (3) statistical-modeling-derived scales (Carter et al., 2017).

**Unassisted clinical prediction.** Unassisted clinician prediction refers to the generation of suicide assessment wherein the clinician would conduct interviews with the patient and ask about symptoms of conditions such as mood disorders, substance use, and trauma without the assistance of other psychometric measures. Evidence has found that these interviews in and of themselves can have an ameliorating effect in patients who may be at risk of suicide by simply giving patients the opportunity to talk about what they are experiencing (Dazzi, Gribble, Wessely, & Fear, 2014; Blades, Stritzke, Page, & Brown, 2018). The clinical interview also acts as means to build rapport between clinician and patient, strengthening the clinician-patient relationship.

The Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders (DSM), often referred by the acronym SCID, is a widely used interviewing method by clinicians and considered a "gold standard" clinical interview (Drill, Nakash, DeFife, & Westen, 2015). It is also an example of an unassisted clinician prediction method. The SCID was developed in 1987 by Dr. Robert Spitzer and colleagues as a response to the lack of a structured, clinical interview that assessed all mental illnesses described in the DSM-III (Spitzer, Williams, Gibbon, & First, 1992). The SCID provides standardized procedures to assess for and diagnose a majority of mental illnesses in accordance with the DSM-III.

The structure of the SCID is based upon diagnostic algorithms such that multiple diagnostic hypotheses can be tested in tandem (Spitzer et al., 1992). Questions are grouped based on diagnoses and symptoms, and if a required symptom is not present, then the interviewer moves onto the next group of questions. For example, if a patient does not present with manic symptoms (a required criterion for the diagnosis of Bipolar I), then a clinician would skip over other questions pertaining to a Bipolar I diagnosis and move onto questions pertaining to a different disorder. Each group of questions begins with a close-ended "yes" or "no" question, which is followed with an open-ended question such as "what was that like?". These open-ended questions allow for more disclosure of symptoms. Often, clinicians will include their own questions based on their own clinical judgements.

Over the years, the SCID was adapted to include changes made in the newer DSM

editions (e.g., SCID-IV for the DSM-IV). Additionally, variations on the SCID were developed such as the SCID-5-AMPD, which assess the DSM-5 alternative model for personality disorders, and the SCID-5-PD, which evaluates 10 of the personality disorders described in the DSM-5 (First, Skodol, Bender, & Oldham, 2017; First, Williams, Benjamin, & Spitzer, 2016). The SCID has also been demonstrated to have moderate to high reliability and validity, including high inter-rater reliability (Shankman et al., 2018; Drill et al., 2015).

Another well-known and widely used semi-structured clinical interview is the Columbia Suicide Severity Rating Scale (CSSRS) (Posner et al., 2011). The CSSRS specifically asks questions pertaining to suicidal ideation, intent, and plan. If patients indicate suicidal ideation or plan, clinicians are prompted to ask followup questions provided by the CSSRS. Each question asked holds a number weight that is tallied at the end to provide a suicide risk score. The CSSRS has shown high reliability and validity across a variety of samples and settings and is considered a gold standard in suicide assessment by the Food and Drug Administration (FDA) for use in clinical trials (Posner et al., 2011; Food, Administration, et al., 2012). However, there is evidence to suggestion that the CSSRS should no longer be considered a gold standard as it may not fully address all suicide risk factors or ideations (Giddens, Sheehan, & Sheehan, 2014; Meyer et al., 2010).

Unfortunately, the effectiveness and accuracy of interviews rely on the truthfulness of the patient as well as the actions of the clinician. If patients are not truthful about their symptoms, then an interview may not produce helpful feedback to the clinician. Further, a clinician's interviewing style can vary the responsiveness of the patient and the truthfulness of the patients' responses (McCabe, Sterno, Priebe, Barnes, & Byng, 2017). For example, one clinician may directly ask "are you suicidal?" while another may more indirectly ask "have you ever had thoughts of ending your life?" While both questions pertain to a patient's suicide risk, the differences between how the question has been phrased could elicit different answers from the same patient. For these reasons, it is common for clinicians to use scales to help capture any missed or looked over suicide risk factors when conducting a suicide risk assessment.

**Assessment scales.** Psychometric scales are a tool that clinicians utilize to obtain as objective information as possible about a patient's disposition. These scales have evolved over time from being just a standardized set of questions that clinicians use to guide interviews to a comprehensive list of questions for a specific condition based on the symptoms for that condition. Furthermore, these scales undergo rigorous psychometric testing to ensure validity and reliability. Scales can be a valuable tool for clinicians as they can be tailored to assess specific conditions such as suicide risk.

Aaron T. Beck developed several scales that assess for different mental illnesses, two of which that assess for depression and suicidal ideation (Dozois & Covin, 2004; Cochrane-Brink, Lofchy, & Sakinofsky, 2000). The Beck Depression Inventory-II (BDI-II) is a 21-item self-report measure that not only assesses symptoms of depression, but also indicates potential suicidality (Beck, Steer, & Brown, 1996; Wang & Gorenstein, 2013). Within the BDI-II, there is an item that screens for suicidal behaviors, which clinicians may use to prompt for further conversation or assessment of suicide risk. The BDI-II has shown high reliability and validity across an extensive range of samples and settings, meaning that the items on the BDI-II are able to reliably capture symptoms of depression and accurately indicate a patient's current depressive state (Wang & Gorenstein, 2013; Cochrane-Brink et al., 2000).

In addition to the BDI-II, there also exists the Beck Scale for Suicidal Ideation (BSI or BSSI) (Cochrane-Brink et al., 2000; Dozois & Covin, 2004). The BSI is a self-report scale that contains 19-items and 5 screening questions that assess a patient's suicidal ideation, plans, and intent (Beck, Kovacs, & Weissman, 1979). Much like its counterpart, the BDI-II, the BSI displays high reliability and validity (Beck, Steer, & Ranieri, 1988; Cochrane-Brink et al., 2000). The BSI is a useful tool in screening for suicidal ideations, which has been linked to higher risk of suicide (Millner & Nock, 2020).

**Statistical-modeling-derived scale.** A statistical-modeling-derived scale is an assessment scale that has been created based upon a statistical model (Carter et al., 2017). A statistical model is a set of assumptions derived from sample data (McCullagh, 2002).

When using statistical models to create scales, researchers take a sample of a population and find relationships between one or more variables in the sample population to represent a construct. For instance, if a statistical model is used to create a scale that produces a depression score, data from a sample population of depressed individuals would be collected, and the data would be examined to find relationships between selected variables. These relationships would then be used to create the scale. Examples of statistical-modeling-derived scales include the re-ACT Self-Harm rule as well as the Repeated Episodes of Self Harm score (Cooper et al., 2006; Spittal, Pirkis, Miller, Carter, & Studdert, 2014).

The re-ACT Self-Harm rule was developed by researchers using data derived from a cohort sample to screen for self-harm presentations in emergency care patients (Steeg et al., 2012). Binary recursive partitioning was used to examine the data and to extract pertinent variable relationships. The rule had high sensitivity and negative predictive value and identified 83 of the 92 patient suicides.

The Repeated Episodes of Self-Harm score utilized a logistic regression statistical test to identify predictors of self-harm cycle repetition (Spittal et al., 2014). The regression model showed high predictive accuracy of over 70%, as well as high sensitivity. The psychometric properties of the Repeated Episodes of Self-Harm score as well as the re-ACT Self-Harm rule suggest that these tools may be useful in assessing self-harm.

### Early Detection and Intervention

Popenhagen and Qualley (1998) discussed the importance of early detection and intervention to decrease instances of suicide. Early detection can be as simple as identifying suicide risk factors such as those described above. These risk factors can also be assessed using clinical interviews or assessment scales. Unfortunately, in a systematic review conducted by Zalsman et al. (2016), there did not seem to exist a perfect suicide prevention strategy. However, reviews of current strategies show a consensus among most strategies that education and identification of suicide risk factors are important first-steps in identifying individuals at risk of suicide (Zalsman et al., 2016; Mann et al., 2005; Isacsson, 2000;

Gould & Kramer, 2001). Once individuals who are at risk of suicide are identified, proper interventions can be put in place to prevent suicide.

### Limitations to Current Methods of Suicide Assessment

While patient-clinician interviews and the use of psychometric scales are two, well-researched methods of patient assessment, they do have limitations in regards to suicide risk assessment. Firstly, a meta-analysis published by Carter et al. (2017) suggested that validation studies for suicide risk psychometric scales may have higher Type I errors due to high suicide prevalence in study populations. This may indicate that the accuracy of suicide risk scales may not be as high as originally believed. Secondly, patient interviews rely heavily on the patient-clinician relationship, which is built upon trust and relies heavily on patient honesty. Finally, as research into suicide and suicide risk factors continues to grow, it brings into question whether current suicide risk assessment and prediction efforts fully appreciate the complex nature of suicide.

It has been suggested by Franklin et al. (2017) as well as Linthicum et al. (2019) that the use of the traditional research model to predict suicide is a potential reason as to why the predictive ability of suicide risk factors remains low. Most studies seem to isolate risk factors and study them individually or within intuitive and rudimentary groupings (Franklin et al., 2017; J. D. Ribeiro et al., 2018). Essentially, these models oversimplify the complex nature of suicide, which results in prediction models that oversimplify the nature of suicide. The resulting model, then, cannot externalize to new data that may deviate from the overly simplified model (Linthicum et al., 2019). Further, studies also focused on short-term suicide prediction, ignoring how the influence of change effects risk factors (J. D. Ribeiro et al., 2018).

Unfortunately, building a predictive model that examines all potential risk factors as well as accounts for change over time can require the analysis of an incredibly large data set with many nonspecific, unrelated variables. Therefore, a new analytical system needs to be utilized to both parse through the extraneous variables as well as build a model that

not only examines all possible risk factors, but is also able to externalize to new data sets. The prediction and analytical capabilities of machine learning can support the building of this kind of model (J. D. Ribeiro et al., 2018). The use of a machine learning model to aid in suicide risk assessment follows in suit with the current integration of technology in healthcare through the clinical decision support systems (CDSSs).

### Technology in Patient Care

To investigate the potential positive impact of machine learning in the prediction of suicide, it is important to investigate the current use of technology in patient care and how it has historically improved patient outcomes. For example, the adoption of the electronic health record (EHRs) has improved patient outcomes as EHR data can be efficiently processed and used to elucidate patterns in the data. These patterns can aid in prediction when using the right techniques such as machine learning algorithms (Chennamsetty, Chalasani, & Riley, 2015). The use of EHR data has made great impact in the healthcare industry through the advent of patient alert systems and computers that assist in diagnosis.

In response to a 1991 report that said 100,000 deaths in the US medical care system could be attributed to medical errors, a paper was published outlining the positive effect of information technology in healthcare (Donaldson, Corrigan, Kohn, et al., 2000; Bates et al., 2001). The authors reported that, since the implementation of information technology in healthcare, patient outcomes were improved when compared to before these systems were put into place. One of these information technology systems was the creation of the patient alert system.

Perhaps one of the most impactful additions to the medical field was the development of the patient alert system, a system which alerts doctors and nursing staff to the status of a patient based on data provided by the patient (Tate, Gardner, & Weaver, 1990; Rind et al., 1994). Tate et al. (1990) observed that the computerized laboratory alerting system (CLAS) would alert doctors in the anticipation that a patient was in distress based on data such as low white cell count, elevated white cell count, falling hematocrit, and metabolic

acidosis. When researchers conducted a pre-CLAS-post-CLAS study to assess the CLAS's effects on patient outcomes, they found that there was a 12% increase in the proportion of patients who received proper treatment with the CLAS system, and the average amount of time spent in life threatening situations significantly decreased. Additionally, the mean days spent in the hospital fell from 14.6 days to 8.8 days.

Similar results for the patient alert system were found in 1994 by Rind et al. (1994) on the effects of computer-based alert systems. In this study, the system alerted doctors, on average, 21.6 hours earlier to the predicted patient crisis than before the alert system was put into place. This helped to reduce the numbers of serious impairments in patients. Additionally, the alert system changed doctors' behaviors, causing them to be more vigilant towards their patients. Alert systems have had a significantly positive impact on bettering patient outcomes. The addition of machine learning algorithms to the patient alert system can aid doctors in observing the changes or anticipated changes in a patient's status. Further, clinical decision support systems (CDSSs) have also showed promising results in aiding in the diagnosis of other medical conditions.

An example of an CDSS that has greatly improved disease diagnosis is the electrocardiograms (ECG or EKG) (Brailer, Kroch, & Pauly, 1997; Tsai, Fridsma, & Gatti, 2003). The computer assisted test interpretation system (CATI) was created to assist doctors in the reading of electrocardiogram (ECG). In a randomized controlled study, the use of the CATI reduced reading times by 28% (3.8 seconds) and led to significantly lower false positive rates in the diagnosis of heart disease (Brailer et al., 1997). In a similar study conducted by Tsai et al. (2003), participants were randomly assigned to groups in which one used a computer assisted ECG reading software and the other group did not. Similar results were uncovered as those who used the computer software had more correct diagnoses than the group that did not use the computer software.

Furthermore, in 100 trials examined from 1973 to 2004, CDSSs bettered patient outcomes in all trials (Garg et al., 2005). In 10 trials evaluated for their diagnostic system, coronary ischemia diagnoses were improved, decreasing the rate of unnecessary hospital-

ization or coronary care by 15%. The CDSS also improved time to diagnose acute bowel obstructions from 16 hours to 1 hour. In studies involving active health conditions, 5 out of 7 trials reported improved diabetes care, and 5 out of 13 trials improved cardiovascular health outcomes. There was a 64% improvement of diagnosis, preventative care, disease management, drug prescribing and dosing.

In the evolving age of technology in healthcare, machine learning has also been integrated into the prevention, prediction, and diagnosis of diseases. The integration of machine learning into healthcare would act as another CDSS that doctors can utilize to better patient outcomes.

### A Brief Introduction to Machine Learning

Machine learning is a broad umbrella that encompasses many different types of algorithms used for prediction and classification, each algorithm with their own pros and cons (Wiens & Shenoy, 2018). When creating a model, it is typical to analyze and assess many different algorithms to find one with the best classification sensitivity, specificity, and accuracy. In the case of diagnosis, classification algorithms such as Naïve Bayes, decision trees, and random forests are most used because they are able to predict the class, or label, of an output, such as high suicide risk, medium suicide risk, or low suicide risk. It is also possible to predict numerical data using methods such as linear regression. Once an algorithm has been selected and created, the algorithm needs to be trained through testing and validation.

In testing and validating the machine learning model, the most common way is to train the algorithm on a certain percentage of data, for example, 90%, then test the model on the remaining 10% of the data. Though this is a simplification of the complex process of training and testing, the focus of this explanation is to understand what training and testing are. Training is the process in which the model "learns" how to predict the class of an object (i.e., instance) through input variables (i.e., attributes) by analyzing the data and finding patterns among the attributes that correspond to each instance. Testing the model, also known as validating the model, involves giving the model new data (validation

data) to see how it performs in classifying the data.

It is important that during the training/testing phase of a model, the model does not overfit the data. Overfitting refers to when the model is constructed to reach the highest accuracy on just the training data set, decreasing the model's ability to generalize to other data sets. When this occurs, the model is only useful in analyzing the data set in which it was trained and tested and may not achieve the highest accuracy in analyzing a new data set. In medical care, due to the large variety in which data is collected and stored, it is difficult to find a model that generalizes outside of individual institutes. For this reason, a model's goal should be to create an algorithm that generalizes to a range of settings as a whole (Wiens & Shenoy, 2018).

There are, generally speaking, three types of learning: supervised, unsupervised, and reinforcement learning (Linthicum et al., 2019). Supervised learning takes attributes, uses the algorithm to find the most close-fitting class, and outputs a class in which the instance belongs. This type of learning can be used to classify instances or return regressions. For example, in classification, the algorithm may return a "high suicide risk" or "low suicide risk" class. But in regression, the algorithm may return how many days are likely to lead up to suicide.

Unsupervised learning explores the underlying patterns within the data, usually in the form of scatter plots or clusters (Linthicum et al., 2019). Clusters specifically can be used to determine what groups of patients are similar to one another based on shared attributes. This information can be used to form classes. Unsupervised learning can also be used to compress the data in order to reduce noise and the number of extraneous variables in a large data set.

Finally, reinforcement learning is when the algorithm's performance is enhanced based on the environment (Linthicum et al., 2019). Essentially, the algorithm makes decisions one at a time based on the feedback from the result of that decision. A classic example of reinforcement learning is playing chess against an Artificial Intelligence (AI) program. The algorithm makes decisions for their next moves based on the moves of their opponents.

When creating and selecting a machine learning algorithm, a key deciding factor is the data set itself (Kotsiantis, Zaharakis, & Pintelas, 2007). A data set should have informative attributes to ensure the important attributes can be isolated when making a prediction (Kotsiantis et al., 2007). Further, algorithm selection is based on the structure of the data set and the predictions to be made (Kotsiantis et al., 2007). For example, if there are missing data, a decision tree can be used as they are able to handle missing data well and have an intuitive decision-making process. Each node in a tree represents a decision based on one or more attributes, and each leaf is the class for that instance. Decision trees are very popular algorithms to use, however, they can get unruly with complex data sets, and any change in the data can completely alter the structure of the tree (Kotsiantis et al., 2007).

There are many classifier algorithms to choose from when creating a machine learning system, and it is even possible to use a combination of different algorithms (Kotsiantis et al., 2007). The main objective of an algorithm is to extract significant information from the data set with as much classification accuracy as possible (Kononenko, 2001).

### Machine Learning in Healthcare

Machine learning has already made significant impacts in the field of medicine. It can be used as a preventative measure by predicting a patient's diagnosis and introducing early treatment to prevent the outcome or lessen the impact of the outcome. Further, machine learning models can be more powerful over other statistical analyses and patient alert programs because of its ability to handle missing data and be more sensitive to patterns within the data (Ghassemi et al., 2020).

For example, researchers conducted a case study in the role of machine learning to predict pneumonia risk (Caruana et al., 2015). The algorithm had an area under the curve (AUC), a measure of a system's sensitivity, of 0.857, which indicates a good classifier. These results show the ability of machine learning algorithm to predict pneumonia risk in patients. These positive results show promise for machine learning in aiding in the diagnosis

and prediction of certain diseases.

In addition, the clustering ability of machine learning can also be applied to finding patterns within tests and symptoms. For example, neurological differences in those with bipolar disorder and those without can be analyzed by machine learning techniques to find any differences between the two groups' brain scans (Librenza-Garcia et al., 2017). In this meta-analytic study, researchers found that machine learning techniques were able to find patterns in the brain imaging of bipolar disorder individuals to aid in the diagnosis of bipolar disorder and to assess the risk of bipolar disorder.

Further, machine learning can recognize symptomology patterns from traditional psychometric scales that aid in the diagnosis of mental illnesses (Ma et al., 2019). In this study, a machine learning algorithm was used to shorten the 145-item Affective Disorder Evaluation (ADE) to 17 questions, while still maintaining a high diagnostic accuracy of 97.4%, by finding patterns within the data that most strongly indicated a positive diagnosis. In two sister studies, machine learning algorithms were used to cut down the Autism Diagnosed Observation Scale (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) (Wall, Kosmicki, Deluca, Harstad, & Fusaro, 2012; Wall, Dally, Luyster, Jung, & DeLuca, 2012). The cut down ADOS had an accuracy of 87% for differentiating an autism diagnosis while for the ADI-R the accuracy was reported at 78% (Bone et al., 2015). These studies further show the ability of a machine learning algorithm in finding the most optimal set of symptoms to warrant a diagnosis.

Finding patterns in data is a hallmark of machine learning capabilities. For suicide, which is a difficult illness to diagnose due to a vast array of symptoms, machine learning could be used to identify patterns in patient data that reveal important clustering of attributes that indicate suicidality.

Machine learning has the potential to become a useful tool in many aspects of healthcare from aiding and predicting diagnoses and diseases, to reducing the amount of information needed to be collect for a diagnosis. Machine learning algorithms can do this by finding subtle patterns in a noisy data set and training itself on those patterns to achieve the highest

accuracy possible to aid in prediction.

### Potential Limitations to Technology in Patient Care

While technology and machine learning have made positive impacts in healthcare, there exist limitations that prevent these methods from being fully implemented in healthcare. Firstly, there are concerns in regards to potential biases within a machine learning model as well as the risks of over reliance of a model. And, secondly, human factors and ergonomics research expounds upon trust given to a system and the subsequent downfalls should that trust be ill received by the model.

**Biases in machine learning.** It may seem that a machine learning model is without bias because of the perception that a machine cannot have biases. However, there do exist biases in machine learning. In a survey done by Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019), researchers found that a machine learning algorithm can be biased to favor one result over another. This can have significant implications when using a machine learning algorithm to make decisions.

In the review, researchers give the example of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a machine learning algorithm used by judges to assess the risk of an offender committing another crime. Investigation into COMPAS revealed that the system was biased against African-Americans as African-Americans were more likely to score a higher risk of re-offending than Caucasians. Another example of a biased machine learning system was an artificial intelligence (AI) program that was used to judge beauty pageant contestants. Once again, this AI was shown to be biased against darker-skinned contestants. But where do these biases occur in a system? Researchers provided two possible explanations.

Firstly, bias could stem from biases within the data that is used to train and validate the system. Just a few of the many types of biases in data are representation bias, measurements bias, population bias, and sampling bias (Olteanu, Castillo, Diaz, & Kiciman, 2019; Suresh & Guttag, 2019; Mehrabi et al., 2019). These biases, in particular, relate to the topic of

machine learning in predicting suicide risk.

Representation bias results from how one defines and samples a population (Suresh & Guttag, 2019). For example, in a dataset of patients who have committed suicide versus those who have not, if there is an overwhelming number of patients who did not commit suicide, the data are biased against patients who did commit suicide. Thus, the machine learning algorithm may become biased against patients who are at high risk of committing suicide. Population bias is a similar bias whereby an algorithm will be biased against variables that do not appear as frequently as other variables (e.g., if there are more white patients who commit suicide than black patients, the algorithm may be likely to more frequently give white patients a higher suicide risk than black patients) (Olteanu et al., 2019).

Sampling bias occurs when subgroup data are not randomly sampled (Mehrabi et al., 2019). In a database of patients who have committed suicide versus not, it is impossible to randomly assign patients to the "suicide" or "non suicide" groups. Therefore, these machine learning algorithms that are used to predict suicide risk inherently suffer from sampling biases because of the nature of the data used to test and validate the machine.

Measurement bias occurs based on how the factors inputted into the algorithm is measured in order to produce a result (Suresh & Guttag, 2019). For example, if the algorithm were to use scores on the CSSRS as a measure of suicide risk, anyone who scores highly on the CSSRS will be more likely to receive a "high suicide risk" prediction. Conversely, if an algorithm were to use race as a measure of suicide risk, the algorithm would be biased towards patients who belonged to a the race that more frequently had suicides within the dataset.

The second way bias is introduced into a machine learning algorithm is through the decision making process of the algorithm. Specifically within the field of machine learning, "bias" refers to any decision made to choose one generalization (i.e., hypothesis) over another instead of strictly adhering to the generalizations given within a data set (Dietterich & Kong, 1995). Dietterich and Kong (1995) explains two examples of such biases: absolute bias and

relative bias.

Absolute bias is an assumption by the algorithm that the resulting output belongs solely to some "designated set of functions" (Dietterich & Kong, 1995). An example of this would be the assumption that a machine learning model's prediction of "high suicide risk" belongs definitively to a set of Boolean conjunctions. This bias could result in a scenario were "high suicide risk" can only be achieved if every patient variable satisfies the rules made by the Boolean conjunctions.

Relative bias, or search bias, is the assumption that the output is best attained by one set of functions over another (Dietterich & Kong, 1995). The example provided by Dietterich and Kong (1995) refers to the decision tree algorithm. Decision trees are created using a bottom up model where the simplest and most basic set of functions are used first then added upon until the algorithm can correctly classify the training data. Once the algorithm has found this tree, larger and more complicated trees are not constructed. This bias could be a potential limitation in using machine learning algorithms to predict suicide risk as it may fall into the same trap described above where, in an effort to create the simplest model for classification, the algorithm may oversimplify the complex nature of suicide, resulting in a model that does not externalize well (Linthicum et al., 2019).

**Human factors and ergonomics in relation to the usability and trust of machine learning.** In a study conducted by Tsai et al. (2003), researchers found that medical professionals may be led to mistakenly alter their diagnoses in order to match the results given by a CDSS. Additionally, low inter-personnel reliability when interpreting results from a CDSS, such as the CATI, could result in different diagnoses between healthcare providers even though they may be using the same systems and being given the same results (Povyakalo, Alberdi, Strigini, & Ayton, 2013). These systems were designed to reduce the error made in diagnostics and, overall, they achieved just that. However, the introduction of these systems gave rise to a new and different kind of error that is linked to trust in a system. These kinds of errors that result from trust in a system are studied by human factors and ergonomics.

Human factors and ergonomics is the field of study whereby researchers strive to understand the relationship between humans and the elements within a system in order to optimize human system performance (Karwowski, 2006; Presidential, 1996). Systems are also referred to as automation (i.e., use of a computer to complete a task) (Parasuraman & Riley, 1997). From a human factors and ergonomics perspective, poor implementation, usability, and understanding of automation can make it difficult for doctors to be sufficiently trained to utilize the technology to its fullest, as seen in the CDSS and CATI studies described above (Bates et al., 2001; Garg et al., 2005). Ultimately, human factors and ergonomics research describe humans' ability to work with complex, automated systems (Parasuraman, Sheridan, & Wickens, 2008).

Three models within human factors and ergonomics research have been of particular interest in recent years, and all three can contribute to the explanation of reliance and error in a system. First is the situational awareness model (SA), the second is the mental workload model, and the third is trust in automation. The understanding of these models can help to understand human performance in complex systems (Parasuraman et al., 2008).

The SA model refers to the perception of items in the environment, how the items are comprehended, and what the items' statuses will be in the near future (Endsley, 2016). Essentially, SA can be described as how one is aware of their situation and subsequently uses their understanding of the situation to understand what will happen next. The principles of the SA model can be applied to machine learning in this way: knowing what variables are inputted into the algorithm, understanding how those variables and algorithm interact, and how those variables will lead to the algorithm's output. However, it is of importance to note that SA is not a choice or performance of a model, but rather the representation of a continuous diagnosis of an ever changing world (Parasuraman et al., 2008).

In conjunction with SA is the mental workload model. Mental workload is simply the mental resource that is demanded by a task in comparison to the mental resources able to be given by the human worker (Parasuraman et al., 2008). It is hypothesized that workers will tend to rely on automation when mental workload is great (Parasuraman & Riley, 1997).

Mental workload is similar to SA as is it also not a choice or performance of a model, but an assessment of the system's strain on the worker. By assessing the mental workload of a system, one could feasibly refine the system such that the mental workload does not negatively affect the worker. On the topic of machine learning in predicting suicide, the use of the machine learning algorithm should not contribute so much to the clinician's mental workload such that it becomes a hindrance to the clinician.

Trust in automation also contributes greatly to human performance and error in a system. While a system may provide good SA and a balanced mental workload, much of human performance in a system depends on trust. In a study conducted by Lee and Moray (1994), researchers found that when trust in automation exceeded confidence in the ability to do a task by hand, humans turned to automation to get the job done. Conversely, if confidence overtook trust, then humans would take control of the system manually. This phenomenon can also be seen in the CATI and CDSS above. The medical professionals in those studies were shown to have changed their accurate diagnoses to match the diagnoses given by the system, perhaps because their trust in the system exceeded their confidence to reject the system's diagnoses. Another possible explanation could be that the medical professionals had too much trust in the automation, which led to the misuse of the automation as a diagnostic answer rather than a diagnostic aid (Parasuraman & Riley, 1997). Unfortunately, when an error does occur with the use of automation, especially if the error is a costly one, trust in automation decreases (Parasuraman & Riley, 1997). In such a situation, the implementation of the automation can hinder performance if the worker is constantly second guessing the ability of the machine. Understanding and allocating the appropriate trust in automation is crucial for establishing how to best use automation, avoid misuses of automation, and have proper reliance on automation (Parasuraman & Riley, 1997; Parasuraman et al., 2008; Lee & See, 2004).

The potential limitations of using machine learning in the prediction of suicide risk are encapsulated by aspects of human factors and ergonomics as well as potential biases of the machine. In regards to machine learning biases, concerns arise because of the nature of the

data on which the machine is tested and validated as well as the machine's algorithm itself. Many machine learning studies are performed on retrospective data. These data could hold many potential biases that could not only negatively impact the external validity of the machine, but also bias the results given (Nemati et al., 2018; Shimabukuro, Barton, Feldman, Mataraso, & Das, 2017; Olteanu et al., 2019; Suresh & Guttag, 2019; Mehrabi et al., 2019). Within the machine learning algorithm, other biases arise in regards how the machine comes up with its results.

There are also concerns about the effects on SA, mental workload, and trust in the machine learning algorithms. Should the introduction of the algorithm negatively impact SA or mental workload, this could potentially result in human error in operating the machine. Additionally, trust in the machine may be affected by how accurately the machine works and how much confidence the clinician has in their own diagnostic processes. Others have also raised the concern of the "black box" nature of these machine learning systems, as some of systems are not overtly apparent in how they use data from electronic health records (EHRs) to make decisions and diagnoses, which may cause distrust from physicians (Bates et al., 2001).

### Potential impact of machine learning in suicide prediction

Despite the potential limitations in the use of machine learning, previously published studies have shown that machine learning algorithms can predict suicide as far out as 365 days from the event, and as close as 30 days to the event with an area under the curve (AUC) of over .80, compared to traditional methods that have AUCs of .50 (Franklin et al., 2017; Linthicum et al., 2019; Oh, Yun, Hwang, & Chae, 2017; J. Ribeiro et al., 2016; J. D. Ribeiro et al., 2018; Walsh, Ribeiro, & Franklin, 2017, 2018). This AUC can be interpreted as a good level of accuracy for the machine. An overview of multiples studies has also found that machine learning algorithms make more accurate predictions than traditional methods (Linthicum et al., 2019).

Bernert et al. (2020) published a review on artificial intelligence and suicide prevention.

Their findings reported a high accuracy (AUC > .90) on the use of machine learning in predicting suicide. However, researchers were unable to synthesize the data and therefore were not able to perform a meta-analytic test. Therefore, to the knowledge of the PI, there has not been a meta-analysis that has collectively pooled the reported accuracies of suicide prediction via machine learning in published studies.

**Purpose of the Research**

This study sought to understand the effects of machine learning models in predicting suicide within a collection of different research publication through a meta-analytic review and present the results in a clinically useful manner. Additionally, this study investigated potential moderators that may have effected the results of the meta-analysis. It was hypothesized that type of machine learning algorithm used, the type of data used, and the definition of suicide would act as moderating variables. Finally, this research attempted to elucidate the potential usefulness of machine learning in the prediction of suicide, as well as uncover future areas of research and current gaps in the literature.

For the purposes of this study, suicide risk was defined as being at risk of attempting suicide or dying by suicide as distinct from ideation because, though ideation is more common than attempted suicide or death by suicide, the presence of ideation does not always progress to suicide. It was by this definition of suicide that articles were screened based on whether or not they predicted suicide risk. This definition was, in part, based upon Posner, Oquendo, Gould, Stanley, and Davies (2007) whereby the authors validated a categorization scheme for suicidal events that is currently utilized by the Federal Drug Administration (FDA). Based on Posner et al. (2007), suicide death, attempt, and ideation are considered suicidal events whereas self-injurious behavior is considered "indeterminate". As such, events such as self-injurious behavior did not fall under the definition of suicide risk used by this study. Furthermore, even though suicidal ideation was defined as a "suicidal" event by Posner et al. (2007), when viewed as a target for a machine learning algorithm, it did not capture the definition of suicide risk that was intended by the PI, which was risk

of committing suicide. Prediction of suicide risk as defined by presence of suicidal ideation was determined to be more akin to predicting risk of suicidality rather than risk of attempt or death. Therefore, the definition of suicide risk was restricted to risk of death or attempt.

## Methods

A meta-analysis is a statistical test that synthesizes data across multiple studies so that the results from any individual study can be observed in context of all the other studies (Borenstein, Hedges, Higgins, & Rothstein, 2011). In many ways, a meta-analysis can be viewed as a more advanced version of the systematic review. First developed by Gene V. Glass in the 1970s, a meta-analysis pools the standardized mean differences (i.e., differences between the means of two groups such as experimental and control groups) across studies (Glass, 1976). Meta-analytic techniques continued to be refined in order to be applied to other areas of research, such as measurements and assessments, and in the 1980s, Drs. DerSimionian and Laird introduced the methodology to calculate random-effects models, which is part of the model used in the present study (Hasselblad & Hedges, 1995; DerSimonian & Laird, 1986; Schmidt & Hunter, 1977).

This meta-analytic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement (Moher et al., 2009). The PRISMA method provides a standardized framework that allows for efficient methodology in a meta-analytic test. The methodology for this research was based upon the 27-item checklist provided by the PRISMA statement. All articles were housed on EndNote® and Covidence®, and all screening and data extraction were conducted through Covidence®. EndNote® is a tool used to aggregate and organize references and Covidence® is an online tool used to enhance the process of evidence synthesis and data extraction (Babineau, 2014).

### Article Collection and Screening

**Search Strategy.** In accordance with the PRISMA method, a social science reference librarian and an engineering research librarian were consulted for search terms to be

used during the database search phase. Search terms as well as databases searched were: machine learning; machine learning algorithms; prediction; suicid* behavior; suicid* idea*; suicid* thoughts; suicid* self-injury; self-injur*; suicid*. Both psychology and computer science databases were searched: PubMed, PsycINFO, Association for Computing Machinery digital library (ACM DL), Science Direct, and Web of Science. The date range for articles collected was 1992 to February of 2021. The cut-off date of articles used was 1992 to ensure that DSM-IV criteria for any assessment of suicide risk factors were used.

**Inclusionary criteria.** Articles that were included in this analysis needed to contain (a) the use of a minimum of one machine learning algorithm to predict suicide in a patient data set; (b) the name of the algorithm used; (c) the predictors (i.e., attributes) used; (d) sample size, (e) number of positive cases; (f) the resulting accuracy of the algorithm reported as an Area Under the Curve (AUC), sensitivity/specificity, or any other information that could be used to calculate an AUC and, subsequently, a Cohen's $d$ using the formulae provided by Hasselblad and Hedges.

There exist several ways to gauge the accuracy of a machine learning algorithm depending on what feature of the algorithm should be given more weight. For predictive machine learning models, AUCs are the preferred measure of accuracy as the AUC describes how well the model is able to distinguish between classes and focuses on the model's ability to discriminate between positive and negative cases (Hand & Anagnostopoulos, 2013; Bradley, 1997; J. Huang & Ling, 2005).

Furthermore, the published papers were required to be written in English, though translations were allowed, and were required to be peer-reviewed publications. There were no geographic or cultural limitations for the studies included in this meta-analysis.

**Exclusionary criteria.** Publications were excluded if they (a) failed to meet the above mentioned inclusionary criteria and (b) did not predict suicide risk as defined by suicide related death or attempt. Furthermore, studies that only discussed a machine learning algorithm's ability to predict suicide, but did not actually test an algorithm, were excluded.

**Article Screening.** Using the search terms above, a total of 1044 articles were collected from the listed databases. Following the collection of articles, articles then underwent a screening process in Covidence. Firstly, duplicate articles were removed from the database of articles ($n$=194). Once the duplicates were removed, articles were screened based on relevance to the subject via the title of the publication and the abstract. Articles that did not meet relevancy were excluded from the analyses ($n$=586), but were recorded as identified articles.

Next, the remaining articles were assessed based on the exclusionary and inclusionary criteria as listed above. Once again, articles that did not meet these criteria were screened out of the analyses ($n$=194), but were recorded as identified articles. Of the articles that did not meet inclusionary/exclusionary criteria, $n$=56 articles predicted suicidality (i.e., suicidal ideation and self-injury) and $n$=23 predicted suicidal sentiment in social media posts (e.g., expressions of wishing to die). Further, $n$=61 were of the wrong study design (i.e., did not test a machine learning algorithm), $n$=45 either did not predict suicide or did not report one of the accuracy measures as listed above, $n$=5 were abstracts that had subsequently become papers, and $n$=3 were of the wrong "patient" population (e.g., two studies predicted the suicide of Virginia Woolfe using her writings and one study utilized a researcher simulated dataset).

At the ending of screening, there were a total of $n$=70 articles. One ($n$=1) article was excluded during the coding phase due to highly improbable effect sizes reported (e.g., $g$=5.57) as determined by the principal investigator (PI) and a key opinion leader in the field. A final $n$=69 articles were included in the meta-analysis. Figure 1 shows the Covidence generated flow diagram of the screening process in Covidence, not accounting for the study that was excluded post-Covidence screening.

**Data extraction.** The data were coded by three independent coders to ensure accurate data recording. Coder 1 was the PI who coded all articles included in the meta-analysis. Coders 2 and 3 were research assistants affiliated with the Helping Give Away Psychological Science (HGAPS) organization based at the University of North Carolina at Chapel

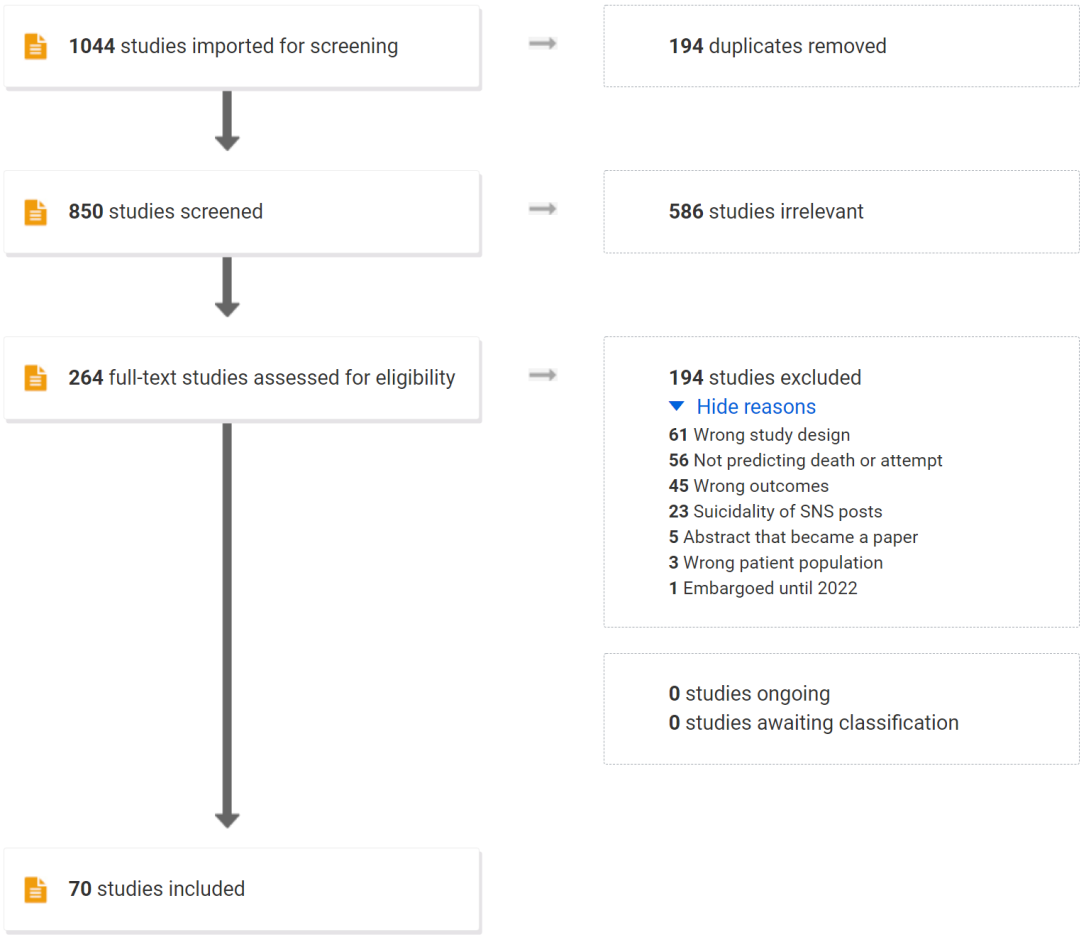


*Figure 1.* PRISMA flow diagram for article selection exported from Covidence. Note that one article was excluded post-Covidence screening, bringing total articles included in analysis to 69.

Hill. Coders 2 and 3 both coded a random selection of 50% of the articles in total. Research assistants were trained on the methodology and theory of meta-analyses by reading methodology papers on how to conduct meta-analyses based on the PRISMA method as well as published meta-analytic review papers on similar topics to the current research as examples. Further, they familiarized themselves with the topics of suicide and machine learning through instruction from the PI as well as reading suicide and machine learning lessons found on the Internet.

Articles for coders 2 and 3 were randomly assigned by giving each article included in the meta-analysis a number and using a random number generator to assign articles to each coder. For every article double coded, the PI conducted a weekly focus group where any queries between the coders' answers were collectively addressed and resolved through consensus by all three parties.

The variables coded in the meta-analysis included the title of the journal in which the article was published, dataset characteristics, accuracy metrics, and the algorithm's characteristics. See Supplemental Materials for the full list of codes used to extract the data. If a study reported results from multiple different models (e.g., if they use different algorithms on the same data set, or they use different predictors in each model), the classification accuracy for all models were recorded as separate instances in the final data and the differences in the algorithms were reported.

In order to streamline the analytic process, type of algorithm used was grouped post-coding into 1 of 13 groupings based on the family in which the algorithm belonged. Regression based algorithms fell under the regression family (e.g., logistic, linear). Methods that used instance-based learning (i.e., learning by comparing new instances to previously encountered instances) such as $k$-nearest neighbor and support vector machines were grouped under the instance based family. All decision tree type algorithms were grouped under decision trees, but random forests and other conglomerate algorithms were categorized as "ensemble". Methods that used Bayesian techniques were grouped under Bayesian, use of artificial neural networks were grouped under ANN, deep learning algorithms were grouped under deep learning, and all algorithms used to transform data into a lower dimensional space such as principal component analyses were grouped under dimensionality reduction. Papers that used only natural language processors (i.e., the natural language processor was not used as a method in a core model) were grouped together, and three studies utilized a unique algorithm the researchers named FALCON. Finally, any paper that utilized a mixed approach of techniques in model construction were classified as Mixed Approach.

**Statistical Methods**

Because meta-analytic techniques are better developed for pooling standardized mean differences rather than AUCs, all AUC values were converted to Hedge's $g$, a standardized mean difference (Hasselblad & Hedges, 1995). Hedge's $g$ was chosen over other effect sizes such as Cohen's $d$ because Hedge's $g$ accounts for the small sample bias that plagues Cohen's $d$ where pooled estimates have an upward bias if the sample size of a study is low (n<21) $d$ (Hedges & Olkin, 2014). Standard formulae given by Hasselblad and Hedges (1995) were used to calculate Hedge's $g$ from AUCs by converting AUCs to Cohen's $d$ first then directly calculating the Cohen's $d$ values to $g$. These formulae were also used to calculate AUCs from sensitivity/specificity values.

A multi-level, mixed-effects model was used in this study. A mixed-effects model accounts for variances that are accounted for by both the fixed effects and random effects models. A fixed effects model assumes that there is one, universal effect size that all studies aim to measure, which causes studies with a greater sample size to be given more weight than smaller studies. In contrast, a random effects model makes no such assumption and takes into account between study heterogeneity. Therefore, in a mixed-effects model, the impact of machine learning on suicide prediction can be measured, like in a fixed effects model, while also accounting for the variance coming from individual differences among the different studies and models, like in a random effects model.

A multi-level approach was taken to account for the experienced nesting due to papers reporting multiple machine learning models. This model not only captures the nesting of effect sizes, but also the within and between study variances while also allowing for a direct comparison of performance measures, negating the need to run separate models for each measure (Konstantopoulos, 2011).

$I^2$ for multi-level meta-analyses as well as sigma values and Cochrane's $Q$ were used to assess the heterogeneity between and within studies to observe whether heterogeneity in effect sizes affected by the variability between the studies and not the data extracted from the studies (Cheung, 2014). Funnel plots and a mixed meta-regression model version

of Egger's test with standardized residuals were used to assess for publication bias and to check for any outliers (Viechtbauer, 2010a). Finally, a meta-regression assessed for the impact of any potential moderating variables. All data analyses were conducted in R version 4.1.0 using the `psych`, `meta`, and `metafor` packages (Viechtbauer, 2010b, 2010a).

## Results

A total of 69 studies published between 1992 and February 2021 contributed 286 Hedge's $g$ effect sizes for the analyses. Table 1 displays the summary of the data for each article included in the analysis. In terms of nesting, one study contributed 24 effect sizes, one study contributed 16, one study contributed 15, two studies contributed 12 each, and the remaining 64 studies contributed between 1 to 10 effect sizes each.

For candidate moderator variables, algorithm family, definition of suicide death and/or attempt, and data source were considered. Tables 2-4 display the summary of sample-level characteristics of effect sizes for each moderator.

Twenty-six of the 69 studies used ensemble machine learning methods, 9 studies used Mixed Approach methods, 6 used decision trees, 6 used regression methods, 5 used artificial neural networks, 5 used deep learning techniques, 5 used instance based methods, 3 used FALCON, 2 used methods listed as other, 1 used Bayesian, and 1 used natural language processors.

25 of the 69 studies utilized the ICD definition of suicide attempt and/or death in their algorithms, 11 studies used a singular survey question (e.g., "have you ever attempted suicide"), 11 studies utilized assessment scales such as the CSSRS, 6 studies used hospital records, 5 studies used interview questions, 3 studies used the NIMH definition, 3 studies used clinical notes, 2 studies used the Study to Assess Risk and Resilience in Servicemembers (STARRS) definition, 1 study used the C-CASA, and 1 study used the ICPC definition.

Finally, for data sources, 13 studies utilized national registries, 11 used data from specialized hospitals (e.g., psychiatric inpatient hospitals), 11 studies used data from the army,

Table 2.
Hedge's $g$ descriptive statistics for moderator algorithm family

| | | Hedge's $g$ | | | |
|---|---|---|---|---|---|
| Algorithm Family | Percent Attribution | Minimum | Maximum | Mean | Std. Dev. |
| ANN | 7.69% | 0.44 | 2.69 | 1.59 | 0.56 |
| Bayesian | 0.70% | 0.75 | 1.88 | 1.32 | 0.79 |
| Decision Tree | 9.46% | 0.32 | 3.06 | 1.26 | 0.6 |
| Deep Learning | 5.94% | 0.48 | 3.19 | 1.94 | 0.74 |
| Dimensionality Reduction | 1.40% | 1.22 | 1.3 | 1.26 | 0.04 |
| Ensemble | 34.62% | 0.5 | 2.88 | 1.61 | 0.52 |
| FALCON | 1.40% | 0.16 | 2.35 | 1.52 | 1.01 |
| Instance based | 8.04% | 0.27 | 2.29 | 1.13 | 0.59 |
| NLP | 0.35% | 1.72 | 1.72 | 1.72 | - |
| Other | 0.70% | 0.72 | 2.2 | 1.46 | 1.05 |
| Regression | 9.44% | 0.11 | 2.46 | 1.31 | 0.54 |
| Mixed Approach | 22.38% | -0.14 | 3.35 | 1.31 | 0.68 |

Table 3.
Hedge's $g$ descriptive statistics for moderator suicide definition

| | | Hedge's $g$ | | | |
|---|---|---|---|---|---|
| Suicide Definition | Percent Attribution | Minimum | Maximum | Mean | Std. Dev. |
| Assessment Scale | 17.83% | 0.11 | 2.35 | 1.13 | 0.47 |
| C-CASA | 0.35% | 1.18 | 1.18 | 1.18 | - |
| Clinical Notes | 1.05% | 0.72 | 1.53 | 1.17 | 0.41 |
| Hospital Records | 6.64% | -0.14 | 2.69 | 0.88 | 0.77 |
| ICD | 46.50% | 0.11 | 3.19 | 1.49 | 0.56 |
| ICPC | 0.35% | 1.29 | 1.29 | 1.29 | - |
| Interview | 7.20% | 1.3 | 2.88 | 1.77 | 0.41 |
| NIMH | 8.74% | 0.39 | 3.06 | 1.61 | 0.49 |
| STARRS | 1.75% | 0.39 | 1.73 | 1 | 0.58 |
| Survey Question | 12.59% | 0.65 | 3.35 | 1.99 | 0.61 |
| Unknown | 0.70% | 0.59 | 0.76 | 0.68 | 0.12 |

Table 4.
Hedge's $g$ descriptive statistics for moderator data source type

| | | Hedge's $g$ | | | |
|---|---|---|---|---|---|
| Data Source Type | Percent Attribution | Minimum | Maximum | Mean | Std. Dev. |
| Anonymous Survey or Interview | 13.65% | 0.11 | 2.29 | 1.37 | 0.43 |
| Army | 8.87% | -0.14 | 2.2 | 0.95 | 0.56 |
| Combination | 5.12% | 0.88 | 1.59 | 1.2 | 0.23 |
| Emergency Room | 1.37% | 1.73 | 1.91 | 1.83 | 0.07 |
| General Hospital | 15.36% | 0.32 | 3.35 | 1.58 | 0.71 |
| Outpatient Clinics | 2.05% | 0.54 | 1.8 | 1.22 | 0.5 |
| Registry | 14.68% | 0.11 | 2.46 | 1.33 | 0.5 |
| SNS | 2.39% | 1.73 | 2.2 | 2 | 0.15 |
| Specialized Hospital | 16.38% | 0.16 | 3.19 | 1.56 | 0.83 |
| University Hospital | 20.14% | 0.31 | 2.88 | 1.65 | 0.53 |

11 used anonymous survey or interview data, 10 studies used university hospital data, 7 used general hospital data, 3 used outpatient clinic data, 1 used combinations of different data sources, 1 used emergency room data, and 1 used social media data.

In addition to the distribution of the above moderator variables throughout the 69 studies, the contribution of effect sizes for each variable also indicated these variables as candidate moderators. Algorithms that utilized the ICD definition of suicide attempt and/or death contributed 45% of all effect sizes ($k$) while assessment scales contributed 17%. Approximately 50% of effect sizes were acquired from two families of machine learning methods: ensemble ($k$=99) and Mixed Approach ($k$=64). Finally, of the 10 data source types, hospital data (general, specialized, university) contributed 50% of the 286 effect sizes, followed by approximately 15% from registries, 13% from anonymous surveys or interviews, 9% from the army, and 11% collectively from combination data sources, emergency room data, social media data, and outpatient clinic data.

Reference standards for the analyses were chosen based on the combination of percent effect size contribution and percent study use with priority given to percent effect size contribution. Because ensemble methods were the most commonly used methods and con-

tributed over 30% of the total effect sizes, it was used as the reference standard in the analyses. Based on the same criteria, the ICD definition of suicide and university hospital data were used as the reference standards. Results included 85,126 positive cases of suicide death or attempt, and a total of 12,752,907 participants.

### Overall Summary of Effect Sizes

The multi-level regression (`rma.mv` function in `metafor`) was used to model the significant nesting of effect sizes in this study by treating the between and within study variances as random effects as described in Konstantopoulos (2011). Cochrane's $Q$ results demonstrated significant heterogeneity of effect sizes ($Q(285\ df) = 66361.51,\ p < 0.0001$). Further, $I^2$ results and $\sigma^{2}$ values showed substantial variance within studies ($\sigma^{2.1} = 0.09$; Level-2 $I^2 = 0.28$) and between study ($\sigma^{2.2} = 0.28$; Level-3 $I^2 = 0.71$), indicating that there are differences among the studies that cannot be attributed to sampling variation alone ($I^2 =$ 0.38, 99.9% ) (Figure 2). The ICC of the true effect size was medium-to-large ($\rho = 0.26$), and the average effect size that was pooled across all studies was large ($g = 1.36$, 95% CI[1.22, 1.49], $p < 0.0001$) (Ahn, Myers, & Jin, 2012). While the large effect size indicated machine learning methods have good discriminative ability in predicting suicide, the large heterogeneity warranted further testing into moderator variables to try and identify the source of variance through the potential moderators (Lipsey & Wilson, 2001).

### Multilevel Meta-Regression Using All Predictors

A fully augmented model that included the chosen moderators (algorithm type, data source type, and suicide risk definition) accounted for a statistically significant amount of variance ($Q(df\ 31) = 50.44,\ p = 0.02$). This augmented model also reduced the random effects within study variance by 0.01 ($\sigma^{2.1} = 0.08$; Level-2 $I^2 = 0.30$), and between study variances by 0.09 ($\sigma^{2.2} = 0.19$; Level-2 $I^2 = 0.70$) (Figure 3). Still, there was significant heterogeneity even within the augmented model ($Q(df\ 254) = 29145.06,\ p < 0.0001$).

The omnibus model indicated that there were significant differences among the sub-

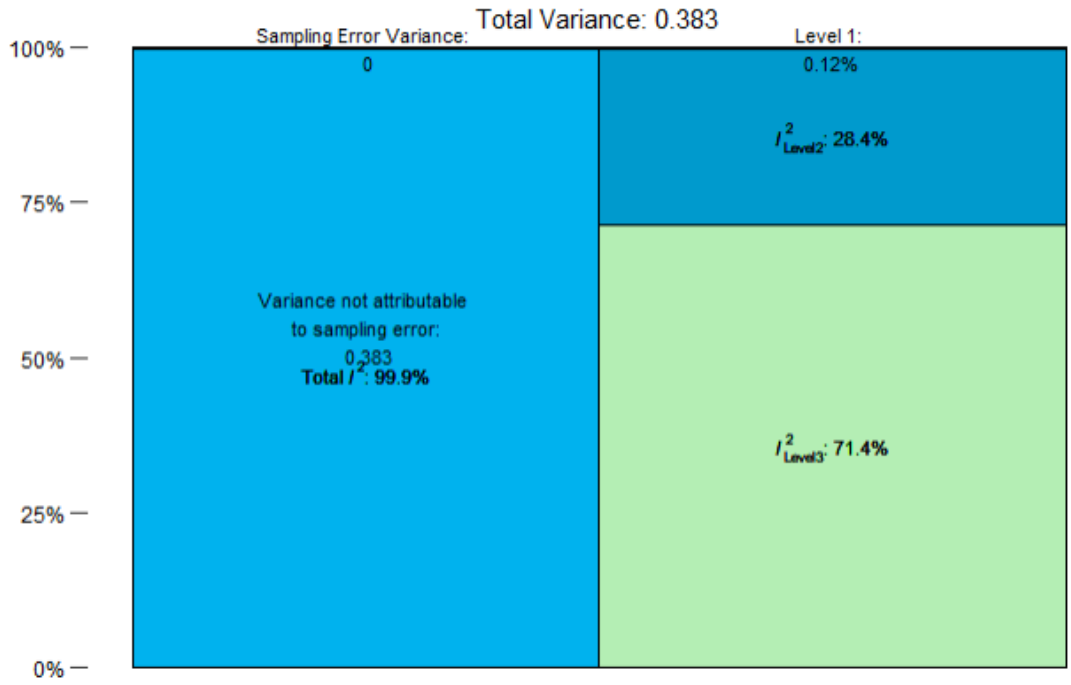


*Figure 2.* $I^2$ for between and within study heterogeneity values for overall model.

groups ($F(31, 254) = 1.63$, $p = 0.02$). The intercept for the fully augmented model was $b =$ 1.56 (95% CI[1.23, 1.90], $p < 0.0001$), indicating a pooled effect size of about $g$=1.56 when including the three moderating variables in the model, and the ICC for the model was large ($\rho = 0.70$).

**Meta-regression for algorithm type.** Analyses were re-run to examine the differences among effect sizes across the different algorithm types. Ensemble methods were used as the reference standard for all the algorithm family types. This model accounted for a statistically significant proportion of the variance ($Q(df\ 11) = 29.68$, $p = 0.002$). The test-of-moderators demonstrated that there was a statistically significant difference between the subgroups of different algorithm types in comparison to ensemble methods ($F(11, 274) =$ 2.70, $p = 0.003$). As seen in Table 5, Bayesian ($B$ = -0.58, $p = 0.015$) and regression ($B$ = -0.37, $p = 0.002$) methods demonstrated statistically significant, lower effect sizes than ensemble methods. Deep learning ($B = 0.33$, $p = 0.016$) reported statistically significant,

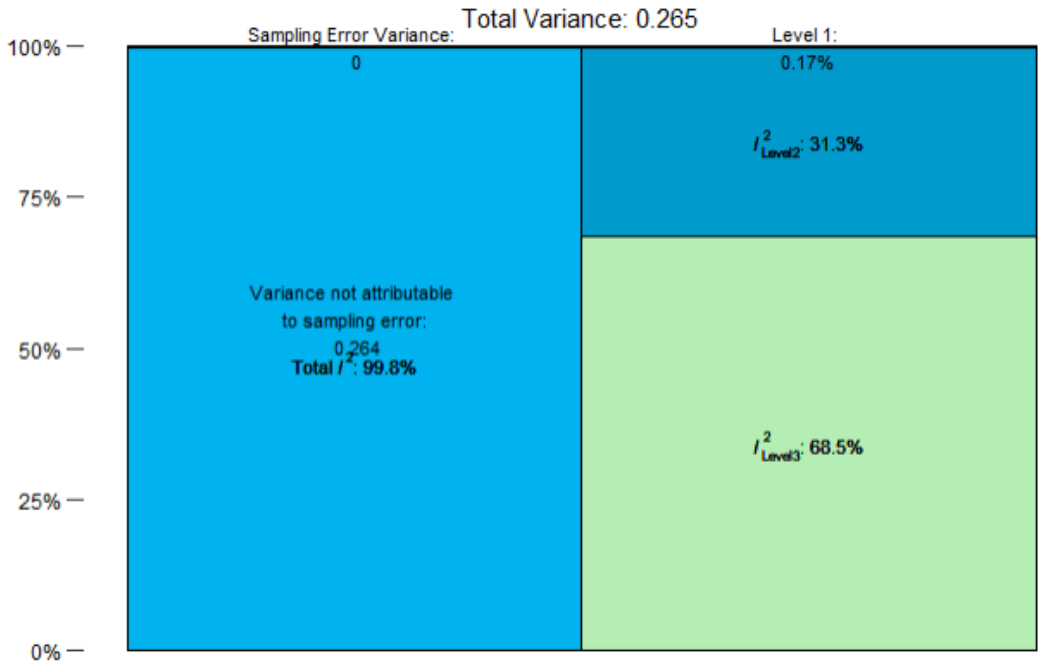


*Figure 3.* $I^2$ for between and within study heterogeneity values for saturated model.

higher effect sizes than ensemble methods.

**Meta-regression for suicide risk definition.** When testing for differences among the different definition of suicide risk, the ICD criteria for suicide death and/or attempt was used as the reference standard for comparison with all other suicide risk definitions. This model did not account for a significant proportion of the variance ($Q(df\ 10) = 15.76$, $p\ 0.11$). Further, the test-of-moderators analysis did not return a significant model ($F(10, 275) = 1.58$, $p = 0.11$). However, individual predictors within the subgroup returned significant results, as seen in Table 6. Survey questions ($B = 0.47$, $p = 0.02$) and interview questions ($B = 0.60$, $p = 0.03$) reported statistically significant, higher effect sizes than ICD.

**Meta-regression for data source type.** To find differences in effect sizes among data source types, university hospital data was used as the reference standard for other data source types. This model did not account for a significant proportion of the variance ($Q(df\ 10) = 10.83$, $p = 0.37$), and the test-of-moderators did not yield a significant model

Table 5.
Regression weights and coefficients for algorithm family meta-regression model

| | $\beta$ | SE | $t$ | df | $p$ | CI [lower, upper] |
|---|---|---|---|---|---|---|
| intercept (Ensemble) | 1.41 | 0.08 | 17.05 | 274 | <.0001*** | [1.25, 1.58] |
| ANN | 0.02 | 0.12 | 0.16 | 274 | 0.87 | [-0.23, 0.27] |
| Bayesian | -0.58 | 0.24 | -2.46 | 274 | 0.01* | [-1.04, -0.12] |
| Decision Tree | -0.16 | 0.1 | -1.53 | 274 | 0.13 | [-0.37, 0.05] |
| Deep Learning | 0.33 | 0.14 | 2.42 | 274 | 0.02* | [0.06, 0.60] |
| Dimensionality Reduction | -0.16 | 0.22 | -0.74 | 274 | 0.46 | [-0.58, 0.26] |
| FALCON | -0.19 | 0.32 | -0.6 | 274 | 0.55 | [-0.81, 0.43] |
| Instance Based | -0.18 | 0.1 | -1.77 | 274 | 0.08 | [-0.37, 0.20] |
| NLP | 0.31 | 0.63 | 0.49 | 274 | 0.63 | [-0.93, 1.54] |
| Other | 0.02 | 0.47 | 0.04 | 274 | 0.97 | [-0.90, 0.94] |
| Regression | -0.37 | 0.12 | -3.11 | 274 | 0.002** | [-0.60, -0.14] |
| Mixed Approach | -0.04 | 0.08 | -0.52 | 274 | 0.6 | [-0.20, 0.12] |

*0.05; **0.01; ***0.0001

Table 6.
Regression weights and coefficients for suicide definition meta-regression

| | $\beta$ | SE | $t$ | df | $p$ | CI [lower, upper] |
|---|---|---|---|---|---|---|
| intercept (ICD) | 1.29 | 0.11 | 12.07 | 275 | <.0001*** | [1.08, 1.50] |
| ICD | -0.61 | 0.56 | -1.1 | 275 | 0.27 | [-1.72, 0.49] |
| Assessment Scale | -0.11 | 0.19 | -0.57 | 275 | 0.57 | [-0.49, 0.27] |
| C-CASA | -0.11 | 0.66 | -0.17 | 275 | 0.86 | [-1.41, 1.18] |
| Clinical Notes | -0.11 | 0.36 | -0.3 | 275 | 0.77 | [-0.82, 0.60] |
| Hospital Records | -0.2 | 0.25 | -0.8 | 275 | 0.43 | [-0.69, 0.29] |
| ICPC | 0.005 | 0.6 | 0.01 | 275 | 0.99 | [-1.17, 1.18] |
| Interview | 0.6 | 0.27 | 2.22 | 275 | 0.03* | [0.07, 1.12] |
| NIMH | 0.14 | 0.33 | 0.41 | 275 | 0.68 | [-0.51, 0.78] |
| STARRS | -0.2 | 0.4 | -0.51 | 275 | 0.61 | [-0.99, 0.58] |
| Survey Question | 0.47 | 0.2 | 2.36 | 275 | 0.02 | [0.08, 0.86] |

*0.05; **0.01; ***0.0001

Table 7.
Regression weights and coefficients for data source meta-regression

| | $\beta$ | SE | $t$ | df | $p$ | CI [lower, upper] |
|---|---|---|---|---|---|---|
| intercept (Uni Hosp) | 1.59 | 0.17 | 9.46 | 276 | <0.0001*** | [1.26, 1.92] |
| Anon Survey or Int | -0.1 | 0.22 | -0.43 | 276 | 0.66 | [-0.54, 0.34] |
| Army | -0.5 | 0.23 | -2.21 | 276 | 0.03* | [-0.94, -0.06] |
| Combo | -0.39 | 0.52 | -0.75 | 276 | 0.46 | [-1.41, 0.63] |
| ER | 0.24 | 0.53 | 0.45 | 276 | 0.66 | [-0.81, 1.29] |
| Gen Hosp | -0.3 | 0.22 | -1.32 | 276 | 0.19 | [-0.74, 0.15] |
| Outpatient | -0.22 | 0.3 | -0.73 | 276 | 0.47 | [-0.82, 0.38] |
| Registry | -0.39 | 0.22 | -1.77 | 276 | 0.08 | [-0.83, 0.04] |
| SNS | 0.4 | 0.52 | -1.77 | 276 | 0.44 | [-0.63, 1.44] |
| Special Hosp | -0.2 | 0.22 | -0.91 | 276 | 0.36 | [0.62, 0.23] |

*0.05; **0.01; ***0.0001

($F(10, 275) = 1.08$, $p = 0.38$) (Table 7). However, data from the army ($B$ = -0.50, p = 0.03) showed significantly lower effect sizes than university hospital effect sizes.

**Publication Biases**

To analyze potential publication biases, funnel plots and a random-effects, mixed-models version of Egger's regression test for publication bias were analyzed (Viechtbauer, 2010a). Figure 4 shows the funnel plot for the fully augmented model. In this model's funnel plot, though effect sizes seem to have considerably variability around the mean, there seems to be no substantial asymmetry. Egger's test results support this as results suggested little evidence of publication bias ($B = 1.09$, $p = 0.33$) and therefore little evidence of asymmetry. Based on this intercept value, it appeared as though larger studies trended towards smaller effect sizes, but not significantly. Further, adjusting for publication biases increased the significance of the findings reported above, suggesting that small sample bias most likely did not contribute to the observed results.

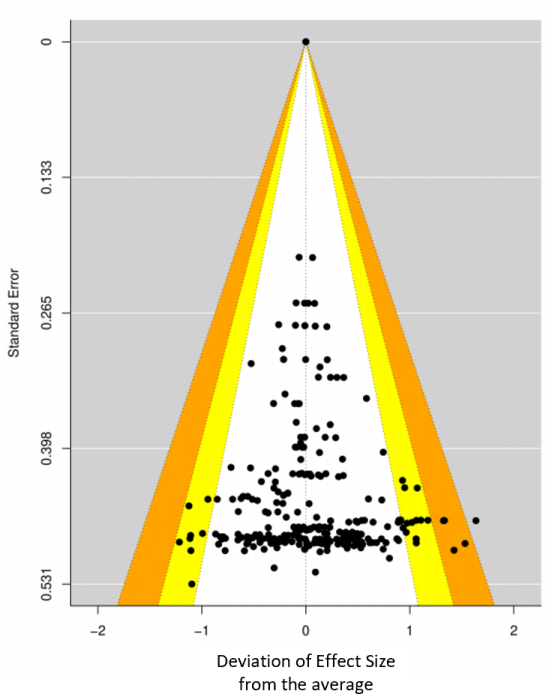


*Figure 4.* Funnel plot of fully augmented model.

## Discussion

The purpose of this meta-analytic review was to assess the utility of machine learning in predicting suicide risk. Based on the results of the random effects, multi-level model, machine learning had a significant effect on the prediction of suicide as defined by suicide death or attempt. The most prominent factor that seemed to play into the discriminative validity of the machine was the choice of machine learning model. The algorithm-family-moderated model returned a statistically significant model whereas the other moderated models did not, indicating that there were significant differences in effect size estimates between the different algorithm types. When converting the estimated, pooled effect sizes for all the models into AUC values, all models showed good to very good discriminative ability. Further, while the other two moderated models were not significant, there were still significant variables within the models that suggested some variables resulted in higher effect sizes than others.

The most frequently reported machine learning algorithms fell within the ensemble methods family with an occurrence of 33.8%. These methods include the random forest algorithms, bagging, and boosting and are often used to tackle complex classification problems such as those dealt with in medicine. Ensemble methods combine several different machine learning techniques into one algorithm in order to decrease variance and bias and increase classification abilities. As a result, it is unsurprising that the label "ensemble methods" encompasses a broad variety of different algorithms and, therefore, were the most commonly used techniques among the publications. This could have potentially contributed to the substantial between-study heterogeneity ($\sigma^{2.2} = 0.29$) found in the multilevel model that was moderated by algorithm families with the reference standard as ensemble methods.

While the test for moderators of suicide definition and data source returned non-significant models, it should be noted that some variables outperformed other variables within these groups. Having the definition of suicide be dictated by anonymous survey or interview questions led to statistically significant higher effect sizes as compared to studies with suicide definitions being decided by the ICD. There exist several possible interpretations of these results. One possible explanation could be that the survey or interview questions provided a definite yes or no classification to the question of "have you attempted suicide" whereas trying to meet the criteria of the ICD's definition of suicide attempt or death may have led to less definite classifications. Essentially, patients may have provided a more precise and accurate response to a direct question about suicide versus trying to ascertain suicide attempt or death in a roundabout way such as going through hospital records postmortem. Another possibility could be that the classifications of suicide attempt or death based on the ICD were not accurate. This could be due to this information being pulled from secondary sources such as electronic medical records that are prone to inaccuracies from factors such as missing information, incorrectly recording patient information, and, in the case of suicide, ambiguous or even untruthful information from patients (Hong, Kaur, Farrokhyar, & Thoma, 2015; Weiner, Wang, Kelly, Sharma, & Schwartz, 2020; McCabe et al., 2017).

Similarly to the meta-regression for suicide definition, the meta-regression for data source type returned an insignificant model. However, within the subgroups, army data source demonstrated significantly lower estimates of effect size when compared to the references standard university hospital data sources. This could be due to a variety of factors such as the lower ratio of positive cases to controls in the army data source studies (1.4% ) in comparison to the studies that used university data (7.1% ). Additionally, it could be the case that data collected by the army regarding a soldier's mental health may be inaccurately or even untruthfully reported due to the stigma and financial and career consequences of disclosing a mental illness diagnosis among military personnel. However, further research is necessary to truly understand why these results were obtained.

### Clinical Interpretability

To provide more clinically meaningful insight into the results of these meta-analyses, the predicted values from the algorithm type meta-regression, the only statistically significant model of all three moderators, were converted into an estimated AUC (Hasselblad & Hedges, 1995). Ensemble methods, the reference standard, have an AUC of approximately 0.84, suggesting good predictive ability.

The one algorithm method that reported statistically significant higher pooled effect sizes was deep learning, which had a reported AUC estimate of 0.89. However, because this AUC approached 0.90, a value that could potentially indicate it is "too good to be true" according to Youngstrom, Salcedo, Frazier, and Perez Algorta (2019), bias analyses were conducted using the deep learning models as a subset. Figure 5 displays the funnel plot for the deep learning algorithms. Based on the plot, it appeared as though there was asymmetry in the reported effect sizes; most studies published reported effect sizes ranging from 1.5 to 3.5. Egger's test of publication bias also returned significant results ($B = 68.2063$, $p < 0.0001$), suggesting that there may be publication biases such that there may be small study effects: smaller studies with large effect sizes may be more frequently published than smaller studies with lower effect sizes.

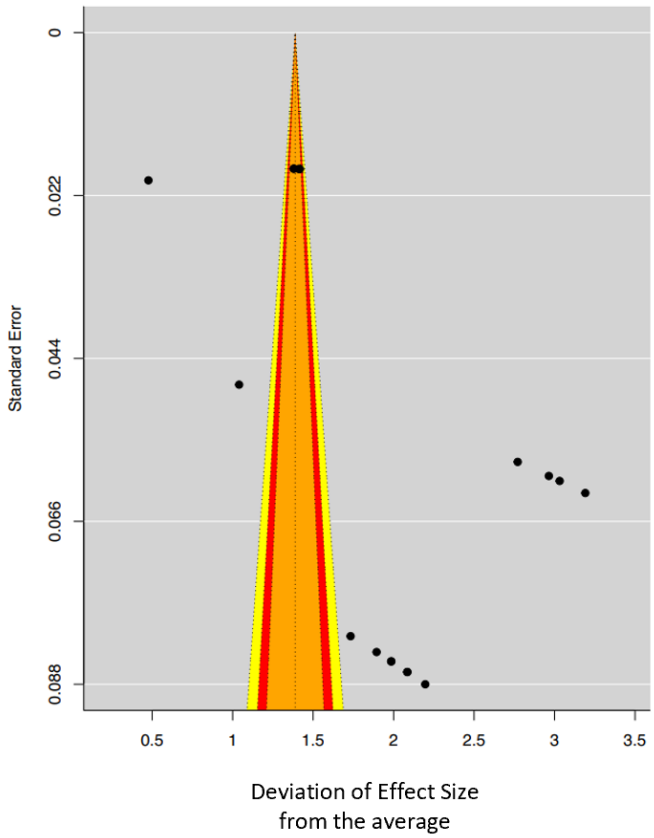


*Figure 5.* Funnel plot of for subgroup Deep Learning algorithm family.

Despite this, most machine learning models had estimated AUCs between 0.80 and 0.89 with Bayesian methods and Mixed Approach methods being the two with lower that 0.80 estimated AUCs. These results suggest that machine learning methods have good discriminative ability in predicting suicide. However, in order to better ascertain this statement, further research must be conducted into why there existed so much variances among the effect sizes within each of the algorithm type groups, especially among studies with low standard error as it is normally believed that as studies approach a standard error of 0, the found effect sizes converge around the mean.

Unfortunately, there was sparse information regarding the human-machine interaction of machine learning and care providers. While no definitive conclusions can be drawn, based on precedent, the introduction of machine learning into a clinical setting will need

to be handled carefully as to maximize the impact of machine learning but minimize the disruptiveness in implementing a new system. Further, as seen in the aforementioned AUC estimates, though machine learning seems to have good predictive ability in predicting suicide risk, caution should be taken in too readily believing such numbers as they may be "too good to be true".

This human factors research may serve as an important, perhaps even critical, component into making a machine learning algorithm a usable clinical tool. As stated previously, current methods of suicide prediction are limited to the simplistic construct that researchers have created about suicide and suicide risk factors. Machine learning is a tool that can be utilized to address this limitation of current methods of suicide assessment. But, in order for machine learning to be implemented as a tool to bridge this gap between capturing a complex model and being a usable tool, usability and human factors research needs to be undertaken. By doing so, not only could suicide prediction become better, but it could also push the evolution of current suicide prediction assessments to be more appreciative of the complexity of suicidality. Additionally, if and when machine learning becomes a usable, clinical tool, it may also help to begin the process of standardizing an assessment plan for suicidality like there exists for assessing other mental illnesses such as depression or anxiety by giving clinicians a launching pad that might further inform their clinical assessments.

### Limitations

Much like how the quality of a machine learning algorithm is heavily dictated by the data it is presented, a meta-analysis can only be as strong as the studies included in the analysis. All studies included were published in academic journals, and the overwhelming majority were in the fields of psychology and medicine. It is uncertain, then, how results reported from fields outside of medicine such as engineering or computer science may impact the results of this analysis. Furthermore, the considerable variances in effect sizes may have implications regarding differences in *what* exactly these studies were examining. It could very well be the case that, while the topics of the studies were about machine learning in

predicting suicide risk, the studies may not all have been assessing the same target diagnosis of suicide risk. This could have been due to a myriad of factors ranging from differences in study design and definition of suicide to an overall, fundamental misunderstanding by the field of how suicide should be viewed.

There is also potential concern about using AUC as a measure of accuracy for machine learning algorithms. AUCs may not be the most effective way in measuring a machine's accuracy in diagnostic settings because in these settings, the data is usually highly class imbalanced (i.e., there are usually many more healthy cases compared to target cases) (Guo, Yin, Dong, Yang, & Zhou, 2008). AUCs account for this imbalance by "glossing over" all potential "bias" (e.g., cut-off scores), but for a model to be useful there needs to be the right "bias" score. To address this, F1 or precision/recall scores could be calculated in addition to the AUC to see how the values compare and to account for any effects due to unbalanced data (Guo et al., 2008).

Additionally, there was high heterogeneity between and within studies despite the presence of three moderators, one of which being statistically significant. While only three moderators were tested, such high level of heterogeneity suggests that variables not included in the analysis contributed to the variance. For example, the calibration of a machine learning algorithm, while standard in practice, can differ greatly depending on the data set, resources available, and expertise of the algorithm coder. Further, the types of predictors and number of predictors as well as the number of training and validation attempts contribute to the overall complexity of the model, which could ultimately result in different performances of models even within the same family. Controlling for these variables could lead to more homogeneous results.

### Future Directions

The results of this meta-analysis demonstrate a potential contribution of the use of machine learning in predicting suicide risk. Future meta-analyses could explore the prediction of suicidality such as predicting presence of suicidal ideations and/or presence of self-injury,

suicidal or non-suicidal. Additionally, future meta-analyses could code for those other variables listed above such as type of predictors, number of predictors, and type of training and calibration. Specifically, it would be interesting to see if a more detailed grouping of algorithms (e.g., grouping by $k$-nearest neighbor models or support vector machine models) would provide a stronger moderator. Searching for more moderators that could better explain and account for the variance in the model will help to elucidate even more fields of research as well as provide better insight into how machine learning is able to achieve high levels of classification accuracy.

In order to further understand the clinical impact of machine learning in suicide prediction, human factors research should also be conducted. Specifically, it would be important to observe the trust or lack thereof from care providers towards the machine especially in reference to the machine's "black box" nature. This relationship between the machine and care provider can then guide the development of protocols to implement these machines as well as how to train providers on the machines. The implications of this human factors research extends beyond just the gathering of knowledge: it provides useful and implementable information to advance the field of suicide prediction.

The tremendous variances in effect sizes warrant further research as well. While it is entirely possible the variances were due to other, unknown factors, one potential source of variance could be that each study was targeting something very closely related to suicide, but not closely related enough to other studies of the same topic. This might indicate that there is more to understand about suicide than is already known that could then produce better classification accuracies.

Finally, the biases of the machines reported in the studies are unknown. Though it is difficult to assess the biases of a machine learning algorithm post-publication, factors such as model complexity and training sequences could provide insight into the likelihood of the machine's biases. Assessing such biases and including them as potential moderators could change the results of the meta-analyses, but in which direction is uncertain.

## Conclusion

Suicide is a devastating act that affects tens of thousands of people yearly. Though there exist evidence-based assessments and scales that are currently being utilized to screen for and predict suicidality, there still exists no gold standard method in assessing suicide. Furthermore, recent research into suicide risk factors have found that the risk factors being used are no better than chance at predicting suicide, perhaps due to the too simplistic way in which researchers have been treating the multi-dimensional features of suicidaility. Though there were limited methods that could fully appreciate such a complex topic, the advancement of machine learning have begun to aid researchers in better understanding and, perhaps, more accurately assessing suicide.

The results of this research found that machine learning indeed has a significant, positive impact on the prediction of suicide and suicide risk as defined by attempted suicide or death by suicide. Further, when converting the summary effect sizes into an AUC for clinical interpretability, AUCs of 0.80 and higher are returned indicating good to very good machine learning discriminative ability. Additionally, although the meta-regressions did not return significant models for data source type or suicide definition, a significant model was returned for algorithm family, which suggested that the type of algorithm chosen had a significant impact on the reported effect sizes. Finally, there was very little evidence of any human factors or usability research on machine learning in predicting suicide risk, highlighting an important area of research that is required in order to address the questions “is machine learning a usable tool in predicting suicide risk".

This review, to the knowledge of the PI, is the first attempt at a meta-analysis on the topic, and results demonstrated consistent findings with previously published literature reviews on the topic that machine learning methods have high accuracies in predicting suicide (Bernert et al., 2020). Further research can be done to expound upon why machine learning has an effect and to what extent can the boundaries of machine learning be pushed not only in suicide classification but in other medical classifications as well.

References

Ahn, S., Myers, N. D., & Jin, Y. (2012). Use of the estimated intraclass correlation for correcting differences in effect size by level. *Behavior research methods*, *44*(2), 490–502.

Apter, A., Bursztein, C., Bertolote, J. M., Fleischmann, A., & Wasserman, D. (2009). Suicide on all the continents in the young. *Oxford textbook of suicidology and suicide prevention*, 621–627.

Babineau, J. (2014). Product review: covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada*, *35*(2), 68–71.

Bates, D. W., Cohen, M., Leape, L. L., Overhage, J. M., Shabot, M. M., & Sheridan, T. (2001). Reducing the frequency of errors in medicine using information technology. *Journal of the American Medical Informatics Association*, *8*(4), 299–308.

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the scale for suicide ideation. *Journal of consulting and clinical psychology*, *47*(2), 343.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–ii. *Psychological Assessment*.

Beck, A. T., Steer, R. A., & Ranieri, W. F. (1988). Scale for suicide ideation: Psychometric properties of a self-report version. *Journal of clinical psychology*, *44*(4), 499–505.

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*, *17*(16), 5929.

Blades, C. A., Stritzke, W. G., Page, A. C., & Brown, J. D. (2018). The benefits and risks of asking research participants about suicide: A meta-analysis of the impact of exposure to suicide-related content. *Clinical psychology review*, *64*, 1–12.

Blair-West, G. W., Cantor, C. H., Mellsop, G. W., & Eyeson-Annan, M. L. (1999). Lifetime suicide risk in major depression: sex and age determinants. *Journal of affective*

*disorders*, *55*(2-3), 171–178.

Bolton, J. M., & Robinson, J. (2010). Population-attributable fractions of axis i and axis ii mental disorders for suicide attempts: findings from a representative sample of the adult, noninstitutionalized us population. *American journal of public health*, *100*(12), 2473–2480.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders*, *45*(5), 1121–1136.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320396001422 doi: https://doi.org/10.1016/S0031-3203(96)00142-2

Brailer, D. J., Kroch, E., & Pauly, M. V. (1997). The impact of computer-assisted test interpretation on physician decision making: the case of electrocardiograms. *Medical decision making*, *17*(1), 80–86.

Carter, G., Milner, A., McGill, K., Pirkis, J., Kapur, N., & Spittal, M. J. (2017). Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *The British Journal of Psychiatry*, *210*(6), 387–395.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*.

CDC. (2019). *Suicide risk and protective factors|suicide|violence prevention|injury center|cdc.gov*. Retrieved 2019-07-15, from https://www.cdc.gov/violenceprevention/suicide/riskprotectivefactors.html

Chennamsetty, H., Chalasani, S., & Riley, D. (2015). *Predictive analytics on electronic health records (ehrs) using hadoop and hive*.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, *19*(2), 211.

Cochrane-Brink, K. A., Lofchy, J. S., & Sakinofsky, I. (2000). Clinical rating scales in suicide risk assessment. *General Hospital Psychiatry*, *22*(6), 445–451.

Cooper, J., Kapur, N., Dunning, J., Guthrie, E., Appleby, L., & Mackway-Jones, K. (2006). A clinical tool for assessing risk after self-harm. *Annals of emergency medicine*, *48*(4), 459–466.

Dazzi, T., Gribble, R., Wessely, S., & Fear, N. T. (2014). Does asking about suicide and related behaviours induce suicidal ideation? what is the evidence? *Psychological Medicine*, *44*(16), 3361–3363.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177–188.

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms* (Tech. Rep.). Technical report, Department of Computer Science, Oregon State University.

Donaldson, M. S., Corrigan, J. M., Kohn, L. T., et al. (2000). *To err is human: building a safer health system* (Vol. 6). National Academies Press.

Dozois, D. J., & Covin, R. (2004). The beck depression inventory-ii (bdi-ii), beck hopelessness scale (bhs), and beck scale for suicide ideation (bss). *Comprehensive handbook of psychological assessment*, *2*, 50–69.

Drill, R., Nakash, O., DeFife, J. A., & Westen, D. (2015). Assessment of clinical information: comparison of the validity of a structured clinical interview (the scid) and the clinical diagnostic interview. *The Journal of nervous and mental disease*, *203*(6), 459.

Endsley, M. R. (2016). *Designing for situation awareness: An approach to user-centered design*. CRC press.

First, M. B., Skodol, A. E., Bender, D. S., & Oldham, J. M. (2017). *User's guide for the structured clinical interview for the dsm-5® alternative model for personality disorders (scid-5-ampd)*. American Psychiatric Pub.

First, M. B., Williams, J. B., Benjamin, L. S., & Spitzer, R. L. (2016). *Scid-5-pd: Structured clinical interview for dsm-5® personality disorders*. American Psychiatric Association Publishing.

Food, Administration, D., et al. (2012). Guidance for industry: suicidal ideation and behavior: prospective assessment of occurrence in clinical trials. *Rockville, MD*.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., & Nock, M. K. (2017). Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. *Psychological Bulletin*, *2*(143), 187-232.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., ... Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, *293*(10), 1223–1238.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, *2020*, 191.

Giddens, J. M., Sheehan, K. H., & Sheehan, D. V. (2014). The columbia-suicide severity rating scale (c–ssrs): Has the "gold standard" become a liability? *Innovations in clinical neuroscience*, *11*(9-10), 66.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8. doi: 10.3102/0013189X005010003

Gould, M. S., & Kramer, R. A. (2001). Youth suicide prevention. *Suicide and life-threatening behavior*, *31*(Supplement to Issue 1), 6–31.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. In *2008 fourth international conference on natural computation* (Vol. 4, p. 192-201). doi: 10.1109/ICNC.2008.871

Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, *34*(5), 492–495.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological bulletin*, *117*(1), 167.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hong, C. J., Kaur, M. N., Farrokhyar, F., & Thoma, A. (2015). Accuracy and completeness of electronic medical records obtained from referring physicians in a hamilton, ontario, plastic surgery practice: a prospective feasibility study. *Plastic Surgery*, *23*(1), 48–50.

Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299–310.

Huang, X., Ribeiro, J. D., Musacchio, K. M., & Franklin, J. C. (2017). Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. *PloS one*, *12*(7), e0180793.

Isacsson, G. (2000). Suicide prevention—a medical breakthrough? *Acta Psychiatrica Scandinavica*, *102*(2), 113–117.

Karwowski, W. (2006). The discipline of ergonomics and human factors. *Handbook of human factors and ergonomics*, *3*.

Kassebaum, N., Kyu, H. H., Zoeckler, L., Olsen, H. E., Thomas, K., Pinho, C., ... others (2017). Child and adolescent health from 1990 to 2015: findings from the global burden of diseases, injuries, and risk factors 2015 study. *JAMA pediatrics*, *171*(6), 573–592.

Kolves, K., & De Leo, D. (2014). Suicide rates in children aged 10–14 years worldwide: changes in the past two decades. *The British Journal of Psychiatry*, *205*(4), 283–285.

Kõlves, K., & De Leo, D. (2016). Adolescent suicide rates between 1990 and 2009: Analysis of age group 15–19 years worldwide. *Journal of Adolescent Health*, *58*(1), 69–77.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89–109.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A

review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3–24.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk. *Archives of general psychiatry*, *54*(4), 337–343.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153–184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Librenza-Garcia, D., Kotzian, B. J., Yang, J., Mwangi, B., Cao, B., Lima, L. N. P., ... Passos, I. C. (2017). The impact of machine learning techniques in the study of bipolar disorder: a systematic review. *Neuroscience & Biobehavioral Reviews*, *80*, 538–554.

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral sciences & the law*, *37*(3), 214–222.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* SAGE publications, Inc.

Luo, F., Florence, C. S., Quispe-Agnoli, M., Ouyang, L., & Crosby, A. E. (2011). Impact of business cycles on us suicide rates, 1928–2007. *American journal of public health*, *101*(6), 1139–1146.

Ma, Y., Ji, J., Huang, Y., Gao, H., Li, Z., Dong, W., ... others (2019). Implementing machine learning in bipolar diagnosis in china. *Translational psychiatry*, *9*(1), 1–7.

Mann, J. J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., Haas, A., ... others (2005). Suicide prevention strategies: a systematic review. *Jama*, *294*(16), 2064–2074.

McCabe, R., Sterno, I., Priebe, S., Barnes, R., & Byng, R. (2017). How do healthcare professionals interview patients to assess suicide risk? *BMC psychiatry*, *17*(1), 1–10.

McCullagh, P. (2002). What is a statistical model? *Annals of statistics*, 1225–1267.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Meltzer, H. Y. (2002). Suicidality in schizophrenia: a review of the evidence for risk factors and treatment options. *Current psychiatry reports*, *4*(4), 279–283.

Meyer, R. E., Salzman, C., Youngstrom, E. A., Clayton, P. J., Goodwin, F. K., Mann, J. J., ... Greden, J. F. (2010). Suicidality and risk of suicide—definition, drug safety concerns, and a necessary target for drug development: a consensus statement. *The Journal of clinical psychiatry*, *71*(8), 0–0.

Millner, A. J., & Nock, M. K. (2020). Self-injurious thoughts and behaviors. *Assessment of disorders in childhood and adolescence*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS med*, *6*(7), e1000097.

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, *46*(4), 547.

NIMH. (2019). *Mih » suicide. nimh.nih.gov.* Retrieved 2019-07-15, from `https://www.nimh.nih.gov/health/statistics/suicide.shtml`

Oh, J., Yun, K., Hwang, J. H., & Chae, J. H. (2017). Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. *Frontiers in psychiatry*, *8*(192).

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making*, *2*(2), 140–160.

Popenhagen, M. P., & Qualley, R. M. (1998). Adolescent suicide: Detection, intervention, and prevention. *Professional school counseling*, *1*(4), 30–36.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., ... others (2011). The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, *168*(12), 1266–1277.

Posner, K., Oquendo, M. A., Gould, M., Stanley, B., & Davies, M. (2007). Columbia classification algorithm of suicide assessment (c-casa): classification of suicidal events in the fda's pediatric suicidal risk analysis of antidepressants. *American journal of psychiatry*, *164*(7), 1035–1043.

Povyakalo, A. A., Alberdi, E., Strigini, L., & Ayton, P. (2013). How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Medical Decision Making*, *33*(1), 98–107.

Presidential, H. W. H. H. (1996). Good ergonomics is good economics..

Ribeiro, J., Franklin, J., Fox, K. R., Bentley, K., Kleiman, E. M., Chang, B., & Nock, M. K. (2016). Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychological medicine*, *46*(2), 225–236.

Ribeiro, J. D., Huang, X., Fox, K. R., & Franklin, J. C. (2018). Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. *The British Journal of Psychiatry*, *212*(5), 279–286.

Rind, D. M., Safran, C., Phillips, R. S., Wang, Q., Calkins, D. R., Delbanco, T. L., ... Slack, W. V. (1994). Effect of computer-based alerts on the treatment and outcomes of hospitalized patients. *Archives of Internal Medicine*, *154*(13), 1511–1517.

Rossen, L. M., Hedegaard, H., Khan, D., & Warner, M. (2018). County-level trends in suicide rates in the us, 2005–2015. *American journal of preventive medicine*, *55*(1), 72–79.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529.

Scott, L. N., Pilkonis, P. A., Hipwell, A. E., Keenan, K., & Stepp, S. D. (2015). Non-suicidal

self-injury and suicidal ideation as predictors of suicide attempts in adolescent girls: A multi-wave prospective study. *Comprehensive psychiatry*, *58*, 1–10.

Shankman, S. A., Funkhouser, C. J., Klein, D. N., Davila, J., Lerner, D., & Hee, D. (2018). Reliability and validity of severity dimensions of psychopathology assessed using the structured clinical interview for dsm-5 (scid). *International journal of methods in psychiatric research*, *27*(1), e1590.

Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ open respiratory research*, *4*(1).

Shneidman, E. S. (1993). Commentary: Suicide as psychache.

Soole, R., Kõlves, K., & De Leo, D. (2014). Factors related to childhood suicides: Analysis of the queensland child death register. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *35*(5), 292.

Spittal, M. J., Pirkis, J., Miller, M., Carter, G., & Studdert, D. M. (2014). The repeated episodes of self-harm (resh) score: A tool for predicting risk of future episodes of self-harm by hospital patients. *Journal of affective disorders*, *161*, 36–42.

Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The structured clinical interview for dsm-iii-r (scid): I: history, rationale, and description. *Archives of general psychiatry*, *49*(8), 624–629.

Steeg, S., Kapur, N., Webb, R., Applegate, E., Stewart, S., Hawton, K., ... Cooper, J. (2012). The development of a population-level clinical screening tool for self-harm repetition and suicide: the react self-harm rule. *Psychological medicine*, *42*(11), 2383–2394.

Stone, D. M., Simon, T. R., Fowler, K. A., Kegler, S. R., Yuan, K., Holland, K. M., ... Crosby, A. E. (2018). Vital signs: trends in state suicide rates—united states, 1999–2016 and circumstances contributing to suicide—27 states, 2015. *Morbidity and Mortality Weekly Report*, *67*(22), 617.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Tate, K. E., Gardner, R., & Weaver, L. (1990). A computerized laboratory alerting system. *MD computing: computers in medical practice*, *7*(5), 296–301.

Tsai, T. L., Fridsma, D. B., & Gatti, G. (2003). Computer decision support as a source of interpretation error: the case of electrocardiograms. *Journal of the American Medical Informatics Association*, *10*(5), 478–483.

Viechtbauer, W. (2010a). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, *36*(3), 1–48.

Viechtbauer, W. (2010b). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software, Articles*, *36*(3), 1–48. Retrieved from `https://www.jstatsoft.org/v036/i03` doi: 10.18637/jss.v036.i03

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS one*, *7*(8), e43855.

Wall, D. P., Kosmicki, J., Deluca, T., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry*, *2*(4), e100–e100.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *3*(5), 457-469.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of child psychology and psychiatry*, *12*(59), 1261-1270.

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the beck depression inventory-ii: a comprehensive review. *Brazilian Journal of Psychiatry*, *35*(4), 416–431.

Weiner, S. J., Wang, S., Kelly, B., Sharma, G., & Schwartz, A. (2020). How accurate is the medical record? a comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *Journal of the American*

*Medical Informatics Association*, *27*(5), 770–775.

Wenzel, A., Berchick, E. R., Tenhave, T., Halberstadt, S., Brown, G. K., & Beck, A. T. (2011). Predictors of suicide relative to other deaths in patients with suicide attempts and suicide ideation: a 30-year prospective study. *Journal of affective disorders*, *132*(3), 375–382.

WHO. (2014). *Preventing suicide a global initiative.*

WHO. (2017). *World health organization.* Author.

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153.

Wilkinson, P., Kelvin, R., Roberts, C., Dubicka, B., & Goodyer, I. (2011). Clinical and psychosocial predictors of suicide attempts and nonsuicidal self-injury in the adolescent depression antidepressants and psychotherapy trial (adapt). *American journal of psychiatry*, *168*(5), 495–501.

Youngstrom, E. A., Salcedo, S., Frazier, T. W., & Perez Algorta, G. (2019). Is the finding too good to be true? moving from "more is better" to thinking in terms of simple predictions and credibility. *Journal of Clinical Child & Adolescent Psychology*, *48*(6), 811–824.

Zalsman, G., Hawton, K., Wasserman, D., van Heeringen, K., Arensman, E., Sarchiapone, M., ... others (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, *3*(7), 646–659.

## Appendix

Data summary of articles including effect sizes, AUC values, variance, and number of participants within each study.

Table 1. *Data summary of articles including effect sizes, AUC values, variance, and number of participants within each study.*

| **Study ID** | **Data Srce** | **Algor. Fam.** | **Sui. Def.** | ***n*(Sui)** | ***n*(Tot)** | **AUC** | **Cohen's *d*** | **Hedge's *g*** |
|---|---|---|---|---|---|---|---|---|
| Adamou 2019 | Registry | Ensemble | | 130 | 828 | 0.71 | 0.76 | 0.76 |
| Adamou 2019 | Registry | Mixed Approach | | 130 | 828 | 0.66 | 0.59 | 0.59 |
| Agne 2020 | Spec. Hosp. | Mixed Approach | Survey Q. | 104 | 959 | 0.95 | 2.33 | 2.32 |
| Ben-Ari 2015 | Army | Ensemble | Clin. Notes | 6,510 | 250,401 | 0.86 | 1.53 | 1.53 |
| Bernecker 2019 | Army | Ensemble | ICD | 168 | 3,528 | 0.82 | 1.29 | 1.29 |
| Bernecker 2019 | Army | Regression | ICD | 168 | 3,528 | 0.62 | 0.43 | 0.43 |
| Bhak 2019 | Univ. Hosp. | Ensemble | Interview | 56 | 182 | 0.98 | 2.89 | 2.88 |
| Bhak 2019 | Univ. Hosp. | Ensemble | Interview | 56 | 182 | 0.95 | 2.34 | 2.33 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 1,113 | 13,325 | 0.99 | 3.35 | 3.35 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 1,113 | 12,001 | 0.98 | 3.03 | 3.03 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 1,113 | 25,326 | 0.98 | 2.77 | 2.77 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 1,113 | 12,001 | 0.96 | 2.77 | 2.77 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 608 | 13,325 | 0.97 | 2.72 | 2.72 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 1,113 | 13,325 | 0.97 | 2.64 | 2.64 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 608 | 13,325 | 0.97 | 2.62 | 2.62 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 1,113 | 26,650 | 0.96 | 2.4 | 2.4 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 1,113 | 39,975 | 0.95 | 2.27 | 2.27 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 1,113 | 12,001 | 0.94 | 2.25 | 2.25 |
| Burke 2020 | Gen. Hosp. | Mixed Approach | Survey Q. | 1,113 | 12,001 | 0.94 | 2.25 | 2.25 |
| Burke 2020 | Gen. Hosp. | Ensemble | Survey Q. | 1,113 | 24,002 | 0.94 | 2.16 | 2.16 |
| Carson 2019 | Gen. Hosp. | Ensemble | Survey Q. | 27 | 73 | 0.68 | 0.66 | 0.65 |
| Chen 2020 | Registry | Ensemble | ICD | 9,099 | 126,205 | 0.9 | 1.8 | 1.8 |
| Chen 2020 | Registry | Ensemble | ICD | 18,682 | 126,205 | 0.89 | 1.77 | 1.77 |
| Chen 2020 | Registry | Ensemble | ICD | 9,099 | 126,205 | 0.89 | 1.76 | 1.76 |
| Chen 2020 | Registry | ANN | ICD | 9,099 | 126,205 | 0.89 | 1.75 | 1.75 |
| Chen 2020 | Registry | Mixed Approach | ICD | 9,099 | 126,205 | 0.89 | 1.75 | 1.75 |
| Chen 2020 | Registry | Ensemble | ICD | 9,099 | 126,205 | 0.89 | 1.73 | 1.74 |
| Chen 2020 | Registry | Ensemble | ICD | 18,682 | 126,205 | 0.88 | 1.66 | 1.66 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Chen 2020 | Registry | Ensemble | ICD | 18,682 | 126,205 | 0.87 | 1.62 | 1.62 |
| Chen 2020 | Registry | ANN | ICD | 18,682 | 126,205 | 0.87 | 1.62 | 1.62 |
| Chen 2020 | Registry | Mixed Approach | ICD | 18,682 | 126,205 | 0.87 | 1.61 | 1.61 |
| Cho 2020 | Registry | Ensemble | ICD | 749 | 372,812 | 0.85 | 1.5 | 1.5 |
| Cho 2020 | Registry | Ensemble | ICD | 749 | 372,812 | 0.82 | 1.28 | 1.28 |
| Choi 2018 | Registry | Deep Learning | ICD | 2,546 | 819,951 | 0.63 | 0.48 | 0.48 |
| Choi 2018 | Registry | Instance based | ICD | 2,546 | 819,951 | 0.58 | 0.27 | 0.27 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.94 | 2.2 | 2.2 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.93 | 2.09 | 2.09 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.93 | 2.09 | 2.09 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.92 | 1.99 | 1.99 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.92 | 1.99 | 1.99 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.91 | 1.9 | 1.9 |
| Coppersmith 2018 | SNS | Deep Learning | Survey Q. | 418 | 836 | 0.89 | 1.73 | 1.73 |
| Delgado-Gomez 2011 | Gen. Hosp. | Instance based | NIMH | 345 | 879 | 0.86 | 1.5 | 1.5 |
| Delgado-Gomez 2011 | Gen. Hosp. | Ensemble | NIMH | 345 | 879 | 0.84 | 1.41 | 1.42 |
| Delgado-Gomez 2011 | Gen. Hosp. | Dim. Reduct | NIMH | 345 | 879 | 0.82 | 1.3 | 1.3 |
| Delgado-Gomez 2011 | Gen. Hosp. | Dim. Reduct. | NIMH | 345 | 879 | 0.82 | 1.28 | 1.28 |
| Delgado-Gomez 2011 | Gen. Hosp. | Instance based | NIMH | 345 | 879 | 0.82 | 1.27 | 1.27 |
| Delgado-Gomez 2011 | Gen. Hosp. | Dim. Reduct. | NIMH | 345 | 879 | 0.81 | 1.23 | 1.23 |
| Delgado-Gomez 2011 | Gen. Hosp. | Dim. Reduct. | NIMH | 345 | 879 | 0.85 | 1.22 | 1.22 |
| Delgado-Gomez 2011 | Gen. Hosp. | Ensemble | NIMH | 345 | 879 | 0.79 | 1.12 | 1.12 |
| Delgado-Gomez 2012 | Univ. Hosp. | Mixed Approach | NIMH | 347 | 881 | 0.92 | 2 | 2 |
| Delgado-Gomez 2012 | Univ. Hosp. | Instance based | NIMH | 347 | 881 | 0.9 | 1.82 | 1.82 |
| Delgado-Gomez 2012 | Univ. Hosp. | Mixed Approach | NIMH | 347 | 881 | 0.89 | 1.76 | 1.76 |
| Delgado-Gomez 2012 | Univ. Hosp. | Instance based | NIMH | 347 | 881 | 0.88 | 1.65 | 1.65 |
| Delgado-Gomez 2012 | Univ. Hosp. | Mixed Approach | NIMH | 347 | 881 | 0.88 | 1.63 | 1.63 |
| Delgado-Gomez 2012 | Univ. Hosp. | Decision Tree | NIMH | 347 | 881 | 0.86 | 1.56 | 1.56 |
| Delgado-Gomez 2012 | Univ. Hosp. | Instance based | NIMH | 347 | 881 | 0.87 | 1.56 | 1.56 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Delgado-Gomez 2012 | Univ. Hosp. | Decision Tree | NIMH | 347 | 881 | 0.86 | 1.53 | 1.53 |
| Delgado-Gomez 2012 | Univ. Hosp. | Decision Tree | NIMH | 347 | 881 | 0.84 | 1.39 | 1.39 |
| DelPozo-Banos 2018 | Registry | ANN | NIMH | 2,604 | 52,102 | 0.82 | 1.29 | 1.29 |
| Edgcomb 2020 | Univ. Hosp. | Decision Tree | ICD | 218 | 15,644 | 0.86 | 1.53 | 1.53 |
| Edgcomb 2021 | Univ. Hosp. | Decision Tree | ICD | 109 | 8,408 | 0.73 | 0.87 | 0.87 |
| Edgcomb 2021 | Univ. Hosp. | Decision Tree | ICD | 40,408 | 841,834 | 0.71 | 0.78 | 0.78 |
| Fan 2020 | Univ. Hosp. | Decision Tree | ICD | 205 | 3,168 | 0.95 | 2.33 | 2.32 |
| Fan 2020 | Univ. Hosp. | Instance based | ICD | 205 | 3,168 | 0.87 | 1.61 | 1.61 |
| Fan 2020 | Univ. Hosp. | Ensemble | ICD | 205 | 3,168 | 0.82 | 1.29 | 1.29 |
| Fan 2020 | Univ. Hosp. | Bayesian | ICD | 205 | 3,168 | 0.73 | 0.75 | 0.75 |
| Fan 2020 | Univ. Hosp. | Regression | ICD | 205 | 3,168 | 0.61 | 0.41 | 0.41 |
| Fan 2020 | Univ. Hosp. | Instance based | ICD | 205 | 3,168 | 0.59 | 0.31 | 0.31 |
| Fernandes 2018 | Outp. Clin. | NLP | ICD | 388 | 500 | 0.89 | 1.72 | 1.72 |
| GarciadelaGarza 2021 | Anon. Sur./Int. | Ensemble | Survey Q. | 222 | 34,653 | 0.86 | 1.51 | 1.51 |
| Gradus 2020 | Registry | Ensemble | ICD | 3,951 | 138,543 | 0.88 | 1.67 | 1.67 |
| Gradus 2020 | Registry | Decision Tree | ICD | 3,951 | 138,543 | 0.87 | 1.59 | 1.59 |
| Gradus 2020 | Registry | Ensemble | ICD | 10,152 | 140,743 | 0.8 | 1.19 | 1.19 |
| Gradus 2020 | Registry | Decision Tree | ICD | 10,152 | 140,743 | 0.77 | 1.04 | 1.04 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.96 | 2.46 | 2.46 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.92 | 2 | 2 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.86 | 1.7 | 1.7 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.87 | 1.61 | 1.61 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.87 | 1.58 | 1.57 |
| Green 2019 | Registry | Regression | ICD | 423 | 2,077 | 0.75 | 0.94 | 0.94 |
| Hack 2017 | Gen. Hosp. | Instance based | ICD | 163 | 1,017 | 0.71 | 0.78 | 0.78 |
| Hack 2017 | Gen. Hosp. | Instance based | ICD | 163 | 1,017 | 0.7 | 0.74 | 0.74 |
| Hack 2017 | Gen. Hosp. | Mixed Approach | ICD | 163 | 1,017 | 0.7 | 0.74 | 0.74 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 31 | 2,390 | 0.87 | 1.59 | 1.59 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 31 | 2,390 | 0.86 | 1.53 | 1.53 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 31 | 2,390 | 0.85 | 1.47 | 1.47 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 47 | 2,390 | 0.84 | 1.44 | 1.43 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 31 | 2,390 | 0.83 | 1.35 | 1.35 |
| Haroz 2020 | Combination | Decision Tree | Ass. Scales | 31 | 2,390 | 0.81 | 1.24 | 1.24 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 47 | 2,390 | 0.81 | 1.23 | 1.22 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 20 | 2,390 | 0.79 | 1.16 | 1.16 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 20 | 2,390 | 0.79 | 1.15 | 1.15 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 20 | 2,390 | 0.79 | 1.12 | 1.12 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 47 | 2,390 | 0.76 | 1.02 | 1.01 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 47 | 2,390 | 0.76 | 1.03 | 1.01 |
| Haroz 2020 | Combination | Mixed Approach | Ass. Scales | 20 | 2,390 | 0.74 | 0.92 | 0.92 |
| Haroz 2020 | Combination | Decision Tree | Ass. Scales | 20 | 2,390 | 0.74 | 0.91 | 0.91 |
| Haroz 2020 | Combination | Decision Tree | Ass. Scales | 47 | 2,390 | 0.73 | 0.88 | 0.88 |
| Hettige 2017 | Spec. Hosp. | Mixed Approach | Ass. Scales | 131 | 345 | 0.71 | 0.78 | 0.78 |
| Hettige 2017 | Spec. Hosp. | Mixed Approach | Ass. Scales | 131 | 345 | 0.71 | 0.78 | 0.78 |
| Hettige 2017 | Spec. Hosp. | Instance based | Ass. Scales | 131 | 345 | 0.7 | 0.74 | 0.74 |
| Hettige 2017 | Spec. Hosp. | Ensemble | Ass. Scales | 131 | 345 | 0.67 | 0.62 | 0.62 |
| Hill 2019 | Anon. Sur./Int. | Decision Tree | Survey Q. | 192 | 4,834 | 0.89 | 1.74 | 1.74 |
| Hill 2019 | Anon. Sur./Int. | Decision Tree | Survey Q. | 192 | 4,834 | 0.85 | 1.45 | 1.45 |
| Hong 2021 | Spec. Hosp. | Instance based | C-CASA | 42 | 66 | 0.8 | 1.19 | 1.18 |
| Huang 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 635 | 954 | 0.89 | 1.74 | 1.73 |
| Huang 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 635 | 954 | 0.84 | 1.4 | 1.4 |
| Huang 2020b | Anon. Sur./Int. | Ensemble | Ass. Scales | 633 | 933 | 0.89 | 1.73 | 1.73 |
| Huang 2020b | Anon. Sur./Int. | Ensemble | Ass. Scales | 322 | 885 | 0.89 | 1.73 | 1.73 |
| Huang 2020b | Anon. Sur./Int. | Ensemble | Ass. Scales | 154 | 285 | 0.89 | 1.73 | 1.73 |
| Huang 2020b | Army | Ensemble | Ass. Scales | 755 | 1,584 | 0.87 | 1.59 | 1.59 |
| Huang 2020b | Outp. Clin. | Ensemble | Ass. Scales | 78 | 182 | 0.9 | 1.81 | 1.8 |
| Jung 2019 | Anon. Sur./Int. | Ensemble | Survey Q. | 7,443 | 59,984 | 0.86 | 1.55 | 1.55 |
| Jung 2019 | Anon. Sur./Int. | Ensemble | Survey Q. | 7,443 | 59,984 | 0.85 | 1.51 | 1.55 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Jung 2019 | Anon. Sur./Int. | Instance based | Survey Q. | 7,443 | 59,984 | 0.85 | 1.48 | 1.48 |
| Jung 2019 | Anon. Sur./Int. | Regression | Survey Q. | 7,443 | 59,984 | 0.85 | 1.47 | 1.47 |
| Jung 2019 | Anon. Sur./Int. | ANN | Survey Q. | 7,443 | 59,984 | 0.85 | 1.47 | 1.47 |
| Kessler 2015 | Army | Decision Tree | STARRS | 68 | 40,820 | 0.89 | 1.73 | 1.71 |
| Kessler 2015 | Army | Mixed Approach | STARRS | 68 | 40,820 | 0.85 | 1.47 | 1.47 |
| Kessler 2017 | Army | Mixed Approach | STARRS | 66 | 234 | 0.72 | 0.82 | 0.82 |
| Kessler 2017 | Army | Mixed Approach | STARRS | 66 | 234 | 0.66 | 0.58 | 0.58 |
| Kessler 2017 | Army | Mixed Approach | STARRS | 66 | 234 | 0.61 | 0.4 | 0.39 |
| Kessler 2020 | Army | Ensemble | ICD | 1,195 | 391,018 | 0.82 | 1.29 | 1.29 |
| Kessler 2020 | Army | Ensemble | ICD | 1,195 | 391,018 | 0.8 | 1.19 | 1.19 |
| Kessler 2020 | Army | Ensemble | ICD | 1,195 | 391,018 | 0.78 | 1.09 | 1.09 |
| Kessler 2020 | Army | Ensemble | ICD | 1,195 | 391,018 | 0.77 | 1.04 | 1.04 |
| Kessler 2020 | Army | Ensemble | ICD | 1,195 | 391,018 | 0.76 | 1 | 1 |
| Levis 2020 | Army | Mixed Approach | Hosp. Rcrds | 246 | 1,232 | 0.58 | 0.29 | 0.29 |
| Levis 2020 | Army | Mixed Approach | Hosp. Rcrds | 246 | 1,232 | 0.53 | 0.1 | 0.1 |
| Levis 2020 | Army | Mixed Approach | Hosp. Rcrds | 246 | 1,232 | 0.49 | -0.05 | -0.05 |
| Levis 2020 | Army | Mixed Approach | Hosp. Rcrds | 246 | 1,232 | 0.46 | -0.14 | -0.14 |
| Lyu 2019 | ER | ANN | Interview | 659 | 1,318 | 0.91 | 1.91 | 1.91 |
| Lyu 2019 | ER | ANN | Interview | 659 | 1,318 | 0.9 | 1.85 | 1.85 |
| Lyu 2019 | ER | ANN | Interview | 659 | 1,318 | 0.9 | 1.82 | 1.81 |
| Lyu 2019 | ER | ANN | Interview | 659 | 1,318 | 0.89 | 1.73 | 1.73 |
| Machado 2021 | Anon. Sur./Int. | Mixed Approach | Interview | 200 | 32,700 | 0.89 | 1.73 | 1.73 |
| Machado 2021 | Anon. Sur./Int. | Ensemble | Interview | 200 | 32,700 | 0.89 | 1.73 | 1.73 |
| Machado 2021 | Anon. Sur./Int. | Mixed Approach | Interview | 200 | 6,350 | 0.89 | 1.73 | 1.73 |
| Machado 2021 | Anon. Sur./Int. | Ensemble | Interview | 200 | 6,350 | 0.89 | 1.74 | 1.73 |
| Machado 2021 | Anon. Sur./Int. | ANN | Interview | 200 | 6,350 | 0.88 | 1.66 | 1.66 |
| Machado 2021 | Anon. Sur./Int. | ANN | Interview | 200 | 32,700 | 0.86 | 1.53 | 1.53 |
| Mann 2008 | Spec. Hosp. | Decision Tree | Ass. Scales | 80 | 457 | 0.8 | 1.19 | 1.19 |
| Mann 2008 | Spec. Hosp. | Decision Tree | Ass. Scales | 118 | 457 | 0.65 | 0.55 | 0.55 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| McKernan 2019 | Univ. Hosp. | Ensemble | ICD | 34 | 8,879 | 0.82 | 1.29 | 1.29 |
| Miche 2020 | Anon. Sur./Int. | Mixed Approach | Interview | 137 | 2,793 | 0.83 | 1.35 | 1.35 |
| Miche 2020 | Anon. Sur./Int. | Ensemble | Interview | 137 | 2,793 | 0.83 | 1.33 | 1.33 |
| Miche 2020 | Anon. Sur./Int. | Regression | Interview | 137 | 2,793 | 0.83 | 1.32 | 1.32 |
| Miche 2020 | Anon. Sur./Int. | Mixed Approach | Interview | 137 | 2,793 | 0.82 | 1.31 | 1.31 |
| Modai 1999 | Spec. Hosp. | ANN | Hosp. Rcrds | 99 | 198 | 0.97 | 2.7 | 2.7 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.92 | 2.01 | 2 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.9 | 1.73 | 1.73 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.88 | 1.58 | 1.57 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.7 | 0.75 | 0.75 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.63 | 0.47 | 0.47 |
| Modai 1998 | Spec. Hosp. | ANN | Hosp. Rcrds | 77 | 161 | 0.62 | 0.44 | 0.44 |
| Modai 2002 | Spec. Hosp. | ANN | Ass. Scales | 50 | 197 | 0.73 | 0.88 | 0.88 |
| Modai 2002 | Spec. Hosp. | FALCON | Ass. Scales | 50 | 197 | 0.54 | 0.16 | 0.16 |
| Modai 2004a | Spec. Hosp. | FALCON | Ass. Scales | 137 | 987 | 0.95 | 2.35 | 2.35 |
| Modai 2004a | Spec. Hosp. | FALCON | Ass. Scales | 137 | 987 | 0.95 | 2.23 | 2.23 |
| Modai 2004b | Outp. Clin. | FALCON | Ass. Scales | 74 | 612 | 0.83 | 1.34 | 1.34 |
| Morales 2017 | Gen. Hosp. | Decision Tree | Hosp. Rcrds | 349 | 707 | 0.74 | 0.89 | 0.89 |
| Morales 2017 | Gen. Hosp. | Decision Tree | Hosp. Rcrds | 349 | 707 | 0.69 | 0.7 | 0.7 |
| Morales 2017 | Gen. Hosp. | Decision Tree | Hosp. Rcrds | 349 | 707 | 0.66 | 0.58 | 0.58 |
| Morales 2017 | Gen. Hosp. | Decision Tree | Hosp. Rcrds | 349 | 707 | 0.59 | 0.32 | 0.32 |
| Obeid 2020 | Spec. Hosp. | Deep Learning | ICD | 835 | 2,505 | 0.99 | 3.19 | 3.19 |
| Obeid 2020 | Spec. Hosp. | Deep Learning | ICD | 835 | 2,505 | 0.98 | 3.03 | 3.03 |
| Obeid 2020 | Spec. Hosp. | Deep Learning | ICD | 835 | 2,505 | 0.98 | 2.97 | 2.96 |
| Obeid 2020 | Spec. Hosp. | Deep Learning | ICD | 835 | 2,505 | 0.96 | 2.77 | 2.77 |
| Obeid 2020 | Spec. Hosp. | Ensemble | ICD | 835 | 2,505 | 0.96 | 2.49 | 2.49 |
| Obeid 2020 | Spec. Hosp. | ANN | ICD | 835 | 2,505 | 0.96 | 2.43 | 2.43 |
| Obeid 2020 | Spec. Hosp. | Instance based | ICD | 835 | 2,505 | 0.95 | 2.29 | 2.29 |
| Obeid 2020 | Spec. Hosp. | Bayesian | ICD | 835 | 2,505 | 0.91 | 1.88 | 1.88 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Obeid 2020 | Spec. Hosp. | Decision Tree | ICD | 835 | 2,505 | 0.87 | 1.59 | 1.59 |
| Oh 2017 | Univ. Hosp. | ANN | Survey Q. | 39 | 736 | 0.93 | 2.08 | 2.08 |
| Oh 2017 | Univ. Hosp. | ANN | Survey Q. | 163 | 736 | 0.89 | 1.73 | 1.73 |
| Oh 2017 | Univ. Hosp. | ANN | Survey Q. | 68 | 736 | 0.87 | 1.59 | 1.59 |
| Passos 2016 | Outp. Clin. | Instance based | Hosp. Rcrds | 43 | 144 | 0.77 | 1.04 | 1.04 |
| Passos 2016 | Outp. Clin. | Mixed Approach | Hosp. Rcrds | 43 | 144 | 0.73 | 0.87 | 0.86 |
| Passos 2016 | Outp. Clin. | Instance based | Hosp. Rcrds | 43 | 144 | 0.65 | 0.55 | 0.54 |
| Poulin 2014 | Army | Other | Clin. Notes | 70 | 210 | 0.7 | 0.72 | 0.72 |
| Ribeiro 2019 | Anon. Sur./Int. | Ensemble | Ass. Scales | 50 | 1,021 | 0.84 | 1.04 | 1.4 |
| Ribeiro 2019 | Anon. Sur./Int. | Ensemble | Ass. Scales | 76 | 1,021 | 0.83 | 1.35 | 1.35 |
| Ribeiro 2019 | Anon. Sur./Int. | Ensemble | Ass. Scales | 24 | 1,021 | 0.82 | 1.29 | 1.29 |
| Roglio 2020 | Spec. Hosp. | Ensemble | Survey Q. | 211 | 422 | 0.73 | 0.88 | 0.88 |
| Roglio 2020 | Spec. Hosp. | Ensemble | Survey Q. | 87 | 247 | 0.68 | 0.66 | 0.66 |
| Rosellini 2017 | Army | Regression | ICD | 169 | 21,832 | 0.74 | 0.91 | 0.91 |
| Rozek 2020 | Army | Other | Interview | 26 | 152 | 0.94 | 2.21 | 2.2 |
| Ryu 2019 | Anon. Sur./Int. | Ensemble | Survey Q. | 1,324 | 2,654 | 0.95 | 2.89 | 2.29 |
| Sanderson 2019 | Registry | Deep Learning | ICD | 3,548 | 39,028 | 0.84 | 1.38 | 1.38 |
| Sanderson 2019 | Registry | Regression | ICD | 3,548 | 39,028 | 0.82 | 1.28 | 1.28 |
| Sanderson 2020a | Registry | Ensemble | ICD | 3,548 | 39,028 | 0.85 | 1.46 | 1.46 |
| Sanderson 2020a | Registry | Deep Learning | ICD | 3,548 | 39,028 | 0.84 | 1.42 | 1.42 |
| Sanderson 2020a | Registry | Deep Learning | ICD | 3,548 | 39,028 | 0.84 | 1.41 | 1.41 |
| Sanderson 2020a | Registry | Deep Learning | ICD | 3,548 | 39,028 | 0.84 | 1.38 | 1.38 |
| Sanderson 2020b | Registry | Ensemble | ICD | 269 | 33,694 | 0.88 | 1.65 | 1.65 |
| Sanderson 2020b | Registry | Regression | ICD | 269 | 33,694 | 0.86 | 1.55 | 1.55 |
| Sanderson 2020b | Registry | Regression | ICD | 269 | 33,694 | 0.08 | 1.25 | 1.25 |
| Shen 2020 | Anon. Sur./Int. | Ensemble | Survey Q. | 682 | 4,882 | 0.93 | 2.04 | 2.04 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 24,993 | 9,685,203 | 0.85 | 1.48 | 1.48 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 24,993 | 9,685,203 | 0.84 | 1.39 | 1.39 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 24,993 | 9,685,203 | 0.84 | 1.39 | 1.39 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 24,993 | 9,685,203 | 0.83 | 1.38 | 1.38 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 1,301 | 9,685,203 | 0.83 | 1.37 | 1.37 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 1,301 | 9,685,203 | 0.82 | 1.31 | 1.31 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 1,301 | 9,685,203 | 0.82 | 1.3 | 1.3 |
| Simon 2019 | Gen. Hosp. | Mixed Approach | ICD | 1,301 | 9,685,203 | 0.82 | 1.29 | 1.29 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 2,383 | 10,275,853 | 0.86 | 1.53 | 1.53 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 2,383 | 10,275,853 | 0.86 | 1.53 | 1.53 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 2,383 | 10,275,853 | 0.86 | 1.52 | 1.52 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 63,805 | 10,275,853 | 0.85 | 1.47 | 1.47 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 63,805 | 10,275,853 | 0.84 | 1.43 | 1.43 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 63,805 | 10,275,853 | 0.84 | 1.43 | 1.43 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 63,805 | 10,275,853 | 0.84 | 1.43 | 1.43 |
| Simon 2019 | Spec. Hosp. | Mixed Approach | ICD | 2,383 | 10,275,853 | 0.84 | 1.38 | 1.38 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 177 | 19,528 | 0.86 | 1.53 | 1.53 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 167 | 17,787 | 0.86 | 1.53 | 1.53 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 114 | 12,566 | 0.86 | 1.53 | 1.53 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 83 | 10,420 | 0.86 | 1.53 | 1.53 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 149 | 16,151 | 0.85 | 1.47 | 1.47 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 139 | 15,064 | 0.85 | 1.47 | 1.47 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 175 | 15,019 | 0.85 | 1.47 | 1.47 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 180 | 41,721 | 0.84 | 1.41 | 1.41 |
| Su 2020 | Gen. Hosp. | Regression | ICD | 60 | 8,366 | 0.81 | 1.24 | 1.24 |
| Tasmim 2020 | Spec. Hosp. | Decision Tree | Ass. Scales | 72 | 189 | 0.72 | 0.84 | 0.84 |
| Tasmim 2020 | Spec. Hosp. | Ensemble | Ass. Scales | 72 | 189 | 0.64 | 0.5 | 0.5 |
| Tasmim 2020 | Spec. Hosp. | Regression | Ass. Scales | 72 | 189 | 0.61 | 0.4 | 0.39 |
| Tiet 2006 | Army | Decision Tree | Clin. Notes | 1,015 | 5,671 | 0.81 | 1.26 | 1.26 |
| vanMens 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 50 | 2,420 | 0.8 | 1.19 | 1.19 |
| vanMens 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 50 | 2,420 | 0.8 | 1.19 | 1.19 |
| vanMens 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 50 | 2,420 | 0.8 | 1.19 | 1.19 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| vanMens 2020a | Anon. Sur./Int. | Ensemble | Ass. Scales | 50 | 2,420 | 0.8 | 1.19 | 1.19 |
| vanMens 2020a | Anon. Sur./Int. | Instance based | Ass. Scales | 50 | 2,420 | 0.79 | 1.14 | 1.14 |
| vanMens 2020a | Anon. Sur./Int. | Decision Tree | Ass. Scales | 50 | 2,420 | 0.77 | 1.04 | 1.04 |
| vanMens 2020a | Anon. Sur./Int. | Regression | Ass. Scales | 50 | 2,420 | 0.72 | 0.82 | 0.82 |
| vanMens 2020a | Anon. Sur./Int. | Instance based | Ass. Scales | 50 | 2,420 | 0.72 | 0.82 | 0.82 |
| vanMens 2020a | Anon. Sur./Int. | Decision Tree | Ass. Scales | 50 | 2,420 | 0.69 | 0.7 | 0.7 |
| vanMens 2020a | Anon. Sur./Int. | Instance based | Ass. Scales | 50 | 2,420 | 0.66 | 0.58 | 0.58 |
| vanMens 2020a | Anon. Sur./Int. | Instance based | Ass. Scales | 50 | 2,420 | 0.63 | 0.47 | 0.47 |
| vanMens 2020a | Anon. Sur./Int. | Regression | Ass. Scales | 50 | 2,420 | 0.53 | 0.1 | 0.1 |
| vanMens 2020b | Registry | Ensemble | ICPC | 574 | 207,882 | 0.82 | 1.29 | 1.29 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 15,945 | 0.92 | 1.98 | 1.99 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 15,945 | 0.86 | 1.53 | 1.53 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.84 | 1.41 | 1.41 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.83 | 1.4 | 1.4 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.83 | 1.4 | 1.4 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.82 | 1.29 | 1.29 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.82 | 1.29 | 1.29 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.82 | 1.29 | 1.29 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.81 | 1.24 | 1.24 |
| Walsh 2017 | Univ. Hosp. | Ensemble | ICD | 3,250 | 5,167 | 0.8 | 1.19 | 1.19 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.97 | 2.66 | 2.66 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.97 | 2.66 | 2.66 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.96 | 2.48 | 2.48 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.96 | 2.48 | 2.48 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.96 | 2.48 | 2.48 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.96 | 2.48 | 2.48 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.95 | 2.33 | 2.33 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 26,056 | 0.94 | 2.2 | 2.2 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.91 | 1.9 | 1.9 |

Continued on next page

**Table 1 – continued from previous page**

| Study ID | Data Srce | Algor. Fam. | Sui. Def. | $n$(Sui) | $n$(Tot) | AUC | Cohen's $d$ | Hedge's $g$ |
|---|---|---|---|---|---|---|---|---|
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.9 | 1.81 | 1.81 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.9 | 1.81 | 1.81 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.9 | 1.81 | 1.81 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.9 | 1.81 | 1.81 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.9 | 1.81 | 1.81 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.89 | 1.73 | 1.73 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 8,034 | 0.87 | 1.59 | 1.59 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.85 | 1.47 | 1.47 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.85 | 1.47 | 1.47 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.83 | 1.35 | 1.35 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.83 | 1.35 | 1.35 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.83 | 1.35 | 1.35 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.83 | 1.35 | 1.35 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.83 | 1.35 | 1.35 |
| Walsh 2018 | Univ. Hosp. | Ensemble | ICD | 975 | 1,471 | 0.82 | 1.29 | 1.29 |
| Zheng 2020 | Registry | Deep Learning | ICD | 448 | 236,347 | 0.77 | 1.04 | 1.04 |
| Zheng 2020 | Registry | Ensemble | ICD | 448 | 236,347 | 0.7 | 0.75 | 0.75 |
| Zheng 2020 | Registry | Mixed Approach | ICD | 448 | 236,347 | 0.06 | 0.38 | 0.38 |
| Zhong 2019 | Registry | Mixed Approach | ICD | 145 | 800 | 0.83 | 1.35 | 1.35 |
| Zhong 2019 | Registry | Mixed Approach | ICD | 145 | 800 | 0.67 | 0.62 | 0.61 |
| Zhong 2019 | Registry | Mixed Approach | ICD | 145 | 800 | 0.53 | 0.11 | 0.11 |
| Zhu 2020a | Univ. Hosp. | Instance based | Hosp. Rcrds | 37 | 90 | 0.92 | 1.98 | 1.98 |
| Zuromski 2020 | Army | Ensemble | ICD | 103 | 7,677 | 0.77 | 1.04 | 1.04 |
| Zuromski 2020 | Army | Mixed Approach | ICD | 103 | 7,677 | 0.76 | 0.98 | 0.98 |
| Zuromski 2020 | Army | Mixed Approach | ICD | 103 | 7,677 | 0.76 | 0.98 | 0.98 |
| Zuromski 2020 | Army | Mixed Approach | ICD | 103 | 7,677 | 0.75 | 0.94 | 0.94 |

## Supplemental Materials

Please view additional supplemental materials here: https://osf.io/8hef6/?view_only=2e0420d53c534da28a7ffd66b82ff9e0