Developing An Expert System for Differentially Diagnosing Borderline Personality Disorder

Shasha Lin

May 2016

A thesis in the Department of Psychology submitted to the faculty of the Graduate School of Arts & Science in partial fulfillment of the requirements for the degree of Master of Arts at New York University

Sponsor: _____
Pascal Wallisch, Ph.D.

Reader:_____
Barry H. Cohen, Ph.D.

Acknowledgments

I would like to first thank Professor Pascal Wallisch, whose generous encouragement and guidance made this project possible. I would also like to thank Professor Barry Cohen, whose passion for statistics and teaching inspired my interest in data analysis, and whose comments and suggestions to this thesis are invaluable. Lastly, I would like to thank the challenges I encountered in the process of writing this thesis, from which I have learned much more than any book or class.

Abstract

Diagnosing mental disorders is a challenging enterprise. Clinical truth is elusive, and data indicate that clinical judgment frequently fail to yield satisfactory accuracy. Bayesian inference enables clinicians to make quantitative assessments based on statistical evidence, and can be easily implemented as a computerized expert system. This paper reviews literature on the accuracy of different diagnostic methods, explains the theoretical foundation for a Bayesian diagnostic system, and provides a minimally viable product (MVP) as a launch pad for the development of future software using Bayesian methods for mental disorder diagnosis.

Chapter One

Clinical Diagnosis versus Statistical Diagnosis

Accurate diagnosis of psychiatric disorders is crucial to effective treatment. Distinct disorders respond to different treatment approaches, and missed diagnoses or misdiagnoses pose major threats to treatment outcomes. The differential diagnosis of borderline personality disorder vs. bipolar disorder is an example that illustrates the foundational importance of accurate diagnosis to effective treatment: several methods of psychotherapy have been shown to be effective in relieving borderline personality disorder (BPD) symptoms, including dialectical behavioral therapy (Linehan et al., 2006), mentalization-based, and supportive psychotherapies (Jørgensen et al., 2013), whereas pharmacotherapy has not been shown to be an effective treatment for BPD pathology as a whole (Francois, Roth, & Klingman, 2015) — generally speaking, drug prescription is not indicated in treating BPD. This is in contrast to the treatment of *bipolar disorder* (BD), where the prescription of mood stabilizers is usually the cornerstone of effective treatment. Failure to differentiate BPD from BD would most likely lead to ineffective, or even harmful treatment plans.

There are two distinct and mutually exclusive methodologies used for evaluating and diagnosing mental disorders: clinical diagnosis versus statistical/model based diagnosis. Clinical diagnoses are reached through inferential judgment by the clinician based on patients' presenting symptoms, family history, holistic impression of patients

from the intake interview, and some psychometric testing results, which can be validated by consensus, or revised for disagreement in case conferences, where several clinicians weigh in their personal judgments. Essentially, this is an entirely experience and intuition based approach: humans who are knowledgeable in the field through experience make the judgment call on a patient's condition.

In comparison, statistical diagnoses follow an explicit input/output mapping, where various modules of patient data are entered into a predictive framework based on the statistical probabilities that each piece of information provides in classifying patients. Diagnoses are based on the statistical results from this framework. Statistical diagnosis is arrived from directly applying an equation (a model) to patient data. Even in cases where clinical ratings are used as input for a statistical model, as long as the clinical judgment is quantified (e.g. in the form of psychometric scores), the method is statistical in nature. On the contrary, if clinical judgment is used in interpreting a statistical diagnosis, and is given the power to revise or override the statistical diagnosis, the method would be clinical in nature, as involvement of subjective judgment renders the diagnosis less than 100% reproducible. To be clear, what matters is what determines the diagnosis on the output level - if the clinician makes the final call, it is clinical diagnosis. If the model has the last say, it is statistical, regardless of inputs. By definition, statistical and clinical diagnoses are two mutually exclusive methods of evaluating mental disorders (Grove, 2005).

Of course, clinical and statistical diagnoses of the same case do not always agree, and practitioners are frequently left to choose between two opposing results. In such

cases where two different diagnoses necessitate completely different courses of treatment, as in the above example of BPD versus BD, clinicians have to decide whether to place their confidence in the statistical model or their own judgment. When faced with such a situation, logic dictates that the clinician's best bet is to choose the diagnostic method that's been shown to consistently demonstrate higher accuracy in clinical settings.

The diagnostic accuracy of statistical versus clinical methods in predicting human health and behavior has been a subject of research for over 60 years (Meehl's book, 1954, on this topic was not the first, but arguably the most influential work that inspired subsequent efforts over later decades to investigate the advantages of one method compared to the other). In one of these classic studies, Goldberg (1965) compared the accuracy of clinicians' judgment and statistical analysis in detecting neurotic versus psychotic patients based on scores on the Minnesota Multiphasic Personality Inventory (MMPI). Out of 7 samples of psychiatric patients totaling 861 MMPI profiles, the statistical rule achieved the same level of accuracy as clinician judgments in 3 samples, and demonstrated modest to substantially higher accuracy in 4 samples, achieving 70% correct decisions, significantly outperforming both the mean accuracy of all clinicians (62%) and the single clinician with highest diagnostic accuracy (67%). It was also found that the 13 clinicians who were experienced in MMPI administration and scoring did not achieve better accuracy than the 16 trainees who were relatively inexperienced with the MMPI. To further examine if abundant training improves clinicians' accuracy, the same 29 clinicians were given 300 new MMPI profiles

to practice accurate diagnosis by having the diagnoses made according to MMPI criteria as immediate feedback. Even after repeated training amounting to 4000 instances of practice diagnosis, none of the clinicians reached the same level of diagnostic accuracy as statistical diagnosis. Parallel to real clinical scenarios where clinicians can choose whether to follow a diagnosis based on statistical algorithm or override the statistical diagnosis with their own judgment, Goldberg conducted a further experiment where a subset of the 29 clinicians were informed of the statistical diagnosis of neurosis versus psychosis, as well as its overall effectiveness, but given the liberty to either adopt this diagnosis or ignore it and make their own judgment. The result suggested that every judge would have done better if they always followed the statistical diagnosis (Dawes, Faust, & Meehl, 1989).

Goldberg's (1970) reanalysis of data collected from the above 29 clinicians sheds light on the reasons behind the inaccuracy of clinical diagnoses. Goldberg developed linear regression models based on the MMPI data and the corresponding diagnoses made by each clinician, reapplied the model to the same MMPI data, and found that the model achieved higher accuracy than each clinician whose judgments were used to develop the model. Using Goldberg's words, *models of man performed better than man himself.* Goldberg attributed this result to the unreliability of human diagnosis: humans have their days: given the same clinical data, they may reach different conclusions depending on arbitrary factors such as mood, energy level, situational and interpersonal distractions, etc. The models developed from clinicians removed this unreliability but kept the general approach of the particular clinician, therefore increasing diagnostic accuracy. This is

further confirmed by the finding that, when data from all 29 clinicians were used to generate a composite model, the model did not achieve higher accuracy than the diagnoses averaged across the 29 clinicians: the relatively large number of clinicians balanced out the arbitrary factors that may notably influence individual clinician judgments.

Several similar studies conducted later confirmed Goldberg's (1965, 1970) findings. Grove, Zald, Lebow, Snitz, & Nelson (2000) performed a meta-analysis of 136 studies comparing statistical and clinical prediction, and found about half of the studies demonstrating the superior accuracy of statistical prediction, while as many showing both statistical and clinical prediction yielding the same accuracy. In only 8 out of 136 studies was clinical prediction more accurate than statistical prediction - consistent with what one would expect from false positive rates - affirming that statistical prediction is "typically as accurate as, or more accurate than clinical prediction". Grove et al. (2000) went further and analyzed if the type of data used had an effect on prediction accuracy; they concluded that the use versus nonuse of behavioral observations, trait ratings, examinee's history data, and psychological tests did not affect the relative accuracy of the two methods of prediction, neither did the amount of data that was made available to clinicians (in the reviewed 136 studies, clinicians either had the same amount or more data as the input in statistical models). The only design variable that significantly influenced the comparison between two methods was whether clinicians were given access to a clinical interview, and counterintuitively, *in cases where a clinical interview*

was available, statistical prediction outperformed clinical prediction by a substantially greater margin.

Grove et al.'s (2000) meta-analysis demonstrated the definitive advantage of statistical diagnosis over clinical diagnosis. It's worth mentioning that they also analyzed different types of predictions investigated in the 136 studies, and found this advantage to be widely verifiable in fields such as general medicine, personality evaluation, mental health diagnosis, and in education and training settings.

Chapter Two

Diagnostic Accuracy: Its Threats and Measurement

In the case of making a differential diagnosis of borderline personality disorder versus other disorders with overlapping symptoms, the disadvantages of clinical diagnosis are glaringly rampant.

Borderline personality disorder (BPD) is a common mental disorder characterized by a pervasive pattern of instability in self-image, affect, and interpersonal relationships. Identifiable symptoms include marked impulsivity that manifests in self-damaging behaviors such as reckless spending and binge eating, pronounced reactivity of mood (episodic dysphoria, irritability, or anxiety), chronic feeling of emptiness, difficulty in controlling anger, unstable interpersonal relationships marked by altering between ideation and devaluation of others, and recurrent suicide or self-harming behaviors (American Psychiatric Association, 2013). The prevalence of BPD is estimated to be between .5% and 5.9% in the general US population, with a median estimate of 1.6% (American Psychiatric Association, 2013). In psychiatric outpatients, the prevalence is

reported to be 10%, and in inpatients, between 15% and 25% —BPD is fairly common within psychiatric populations (Leichsenring, Leibing, Kruse, New, & Leweke, 2011).

BPD is notorious for being frequently misdiagnosed or confused with other mental disorders (Coulston, Tanious, Mulder, Porter, & Malhi, 2012; DeShong & Kurtz, 2013; Eyler, 2007; Fish, 2004; Salzbrenner & Conaway, 2009). One reason for this is the overlapping symptoms between BPD and other common psychiatric disorders, such as bipolar disorder(BD), post-traumatic stress disorder (PTSD), antisocial personality disorder (AsPD), as well as attention-deficit hyperactivity disorder (ADHD). For example, impulsivity is a shared trait of people with BPD, AsPD, and ADHD, mood instability is shared by both BPD, BD and the associated features of PTSD, whereas other PTSD symptoms are similar to core traits of BPD, including cognitive/perceptual disturbance and interpersonal dysfunction.(Zlotnick et al., 2003). It is also common for both BPD and PTSD patients to report a history of sexual abuse. (Heffernan & Cloitre, 2000). Similar clinical presentations of these disorders that are likely to be different in nature obstruct accurate differential diagnosis; as these disorders respond dissimilarly to various treatment strategies, misdiagnosing one for another often leads to ineffective or even harmful treatment. Prescribing mood stabilizers to BPD patients adds unnecessary risks such as lithium poisoning without any benefit, whereas misdiagnosing BPD as PTSD may impair clinicians' strategy in implementing psychotherapies. On the flip side, accurate diagnoses inform most effective treatment decisions: in ADHD patients who have co-occurring BD, prescribing mood stabilizers prior to sequential stimulant and other adjunctive treatments relieves ADHD symptoms 7.5 fold more than prescribing

stimulants that target ADHD symptoms alone (Singh, DelBello, Kowatch, & Strakowski, 2006). Conversely, one might not want to prescribe the kinds of stimulants (typically Ritalin or Adderall) one gives for pure ADHD to someone with BD—there's evidence that stimulant could induce mania or hypomania in BD patients (Wingo & Ghaemi, 2007).

Another challenge to differential diagnosis is the observed high comorbidity between personality disorders and other mental disorders, which has long been a widely acknowledged issue in the study of mental disorders. Studies have found high rates of comorbid PTSD in BPD patients—approximately 30% in a community population (Swartz, Blazer, George, & Winfield, 1990), and 56% in BPD patients seen in psychiatric settings (Zanarini et al., 1998), as well as high comorbidity of BPD in PTSD patients (Shea, Zlotnick, & Weisberg, 1999). Similarly, ADHD has also been found to be overrepresented in BPD population—it's been reported that 38.1% of adult BPD patients had ADHD (Ferrer et al., 2010), in comparison to its much lower prevalence rate in the general adult population (2.5%, Simon, Czobor, & Bálint, 2009). The comorbidity rate between ADHD and BD in children is reported to be even higher: studies have shown that up to 85% of children with a primary diagnosis of BD was also diagnosed with ADHD, while 22% of children with a primary diagnosis of ADHD were also diagnosed with BD (Singh et al., 2006).

There are many explanations for such high comorbidity between BPD and other disorders such as ADHD, BD, PTSD, and AsPD, as well as among these disorders themselves. Singh et al. (2006) investigated four reasons for high comorbidity between

BD and ADHD in pediatric patients enunciated in previous literature: 1) high comorbidity is an artifact of the diagnostic criteria. The current standard for categorizing mental disorders in the United States is the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM 5). DSM 5 is symptom oriented and does not do justice to the complex nature of neurological and etiological factors that underlie mental disorders. It could be the categorization itself that is inaccurate, and therefore many patients diagnosed with multiple disorders may suffer from one unidentified condition, instead of multiple conditions specified by the DSM. 2) ADHD is not a distinct disorder, but rather prodromal to the later development of BD. Namely, ADHD is simply an early manifestation of the onset of BD, and therefore many patients diagnosed with ADHD will continue to develop full-blown BD in later years. 3) ADHD and associated stimulant pharmacotherapy leads to onset of BD. ADHD and BD are two truly independent disorders that are related to each other, in that the symptoms of ADHD and/or its required drug treatments induce a different disorder, BD. 4) ADHD and BD share the same or similar biological risk factors: therefore, patients who have ADHD are also at greater risk of developing BD than the general population. Singh et al. (2006) found some support for the prodromal hypothesis, but no definitive evidence for the other hypotheses.

What Singh et al. neglected to address is another contributing factor for the high comorbidity: many cases of the dual diagnoses may simply be the result of misdiagnoses— ADHD has overlapping manifestations with both mania/hypomania and depression symptoms in BD: distractibility, racing thoughts, impulsivity, and some cognitive deficits are seen in both disorders. Evidence of the misdiagnosis of both

disorders has been shown in the pediatric population. In one such study (Chilakamarri, Filkowski, & Ghaemi, 2011), diagnoses made by consensus between two expert clinicians were compared with diagnoses made by researchers strictly following the DSM-IV, which is the most recent predecessor of DSM 5. (Deeming diagnoses made according to DSM-IV, with the *Structured Clinical Interview for DSM Disorders*, or SCID, as the true condition has been the convention in diagnostic validity research. Since realistically we cannot possibly attain the absolute true condition of patients, i.e. the '*ground truth'* phenomenon. DSM-IV is the closest clinical truth we can attain given the tools we have.) Twenty-nine percent of BD patients as indicated by SCID were misdiagnosed with ADHD by clinicians, and 33% were misdiagnosed with Major Depressive Disorder (MDD). Among those diagnosed by SCID as having MDD, 38% were misdiagnosed with ADHD.

Other studies have confirmed the pervasiveness of misdiagnosis: both missed diagnoses and over-diagnosis have been reported for BPD, BD, PTSD, and AsPD. For example, it's reported that 40% of the clinical population with BPD had a previous misdiagnosis of BD, versus 10% in the clinical population without BPD (this difference is significant, with an odds ratio of 5.0; Ruggero, Zimmerman, Chelminski, & Young, 2010). Besides the kind of over-diagnosis of BD in BPD populations, patients who actually have BD are frequently misdiagnosed with other disorders: 69% of BD patients had been previously misdiagnosed, most frequently with unipolar depression, i.e. MDD (Hirschfeld, Lewis, & Vornik, 2003).

Although comorbidity and misdiagnoses are both inherent challenges to accurate diagnosis, and the latter follows from the former, I have noticed a severe imbalance in the literature that investigates the two respective phenomena: substantially more research effort has been dedicated to reporting the rate of comorbidity, whereas studies of diagnostic validity are few and far between. This is an understandable result of different levels of difficulty involved in the two types of study. Investigating comorbidity is a relatively simple procedure of recording each patient's diagnoses and calculating the percentage of patients with dual diagnoses. In comparison, to study misdiagnoses, researchers have to first make diagnoses with the gold standard, SCID, compare rates of the disorder in question with other disorders, and then compare the gold standard diagnoses with the diagnoses made by clinicians.

Studying the accuracy of a particular diagnostic system also requires a clear understanding of several key statistics that may not be perfectly familiar to clinical researchers: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), as well as the well-known statistic of prevalence. To clarify, sensitivity refers to the percentage of patients who are correctly diagnosed as having a particular disorder out of all who, in fact or per gold standard diagnosis, have the disorder. In parallel, specificity is the percentage of people identified by the diagnostic system as not having a disorder out of people who, in fact, do not have the disorder.

Figure 1 offers an intuitive representation of a diagnostic scenario. All areas included in the grey square represent the whole population. The ellipse to the left denotes people who truly have a particular disorder, e.g. BD, and the ellipse to the right

represents people who are diagnosed as having BD. In this Venn diagram, Sensitivity = $\frac{Area\ 1}{Area\ 1+Area\ 4}=\frac{True\ Positive}{True\ Positive+False\ Negative}$,

specificity$=\frac{Area\ 3}{Area\ 3+Area\ 2}=\frac{True\ Negative}{True\ Negative+False\ Positive}$.

When a diagnostic instrument is developed, sensitivity and specificity are used as standards to evaluate its performance. High sensitivity indicates that the instrument detects the presence of a disorder in most cases; high specificity indicates that the instrument also captures the absence of a disorder in most cases, and for someone who doesn't have a disorder, it's unlikely that their diagnostic result will be positive. When the diagnostic instrument is put to clinical use, however, we evaluate their real performance from the opposite side of the problem: since we don't know the true condition of the patients (*ground truth*[1] problem), but we know their diagnostic result, i.e., whether they meet the criteria for a disorder or not, we care about two things: 1) if

[1] In determining PPV and NPV, researchers are faced with the '*ground truth*' problem: to evaluate a specific diagnostic method, we need to measure its result against the true condition of the patient. Yet how do we determine the 'true' condition? Had we been endowed the ability to know the absolute clinical truth, we would not have needed to develop any diagnostic method in the first place. Spitzer (1983) proposed a method to make the best estimate of *ground truth* by *Longitudinally* assessing patients' symptoms, using *Expert* clinicians to arrive at consensus diagnoses, and using *All Data* available, which means not only patients' own narration, but also information presented by family members, psychiatric ward personnel, any previously seen psychotherapists, etc. (therefore he termed this method '*LEAD*'). LEAD biases towards clinicians in its heavy use of expert clinician judgment, among other flaws. Eventually, most studies of the validity/accuracy of diagnostic instruments simply adopted the Structured Interview of DSM Disorders (SCID) as the gold standard approximation for ground truth, which, incidentally, is developed by Spitzer himself and other collaborators, including First (2002). A few studies approximated the '*LEAD*' standard, and confirmed the relatively superior validity of SCID over usual clinical intake interviews (e.g. Basco et al., 2000, Fenning et al., 1996).

the criteria/clinician indicates that someone has a disorder, what's the probability that they truly have it (PPV); 2) if the criteria/clinician indicates that someone does not have a disorder, what's the probability that they truly do not have it (NPV). Ultimately, we want to achieve an acceptable level of PPV and NPV in clinical settings.

PPV and NPV are determined both by the properties of the specific diagnostic instrument, i.e. sensitivity and specificity, as well as the the prevalence [HAVE YOU DEFINED PREVALENCE YET?] of the disorder in question, and can be calculated with the following equations:

$$PPV=\frac{\text{sensitivity}*\text{prevalence}}{\text{sensitivity}*\text{prevalence}+(1-\text{specificity})*(1-\text{prevalence})}$$

Equation 1.

$$NPV=\frac{\text{specificity}*(1-\text{prevalence})}{(1-\text{sensitivity})*\text{prevalence}+\text{specificity}*(1-\text{prevalence})}$$

Equation 2.

(Zhou, McClish, & Obuchowski, 2009).

What's an acceptable PPV in a primary care setting or a psychiatrist's office? When a patient visits a psychiatrist and is subsequently diagnosed by the psychiatrist with BD, what's the probability that they are misdiagnosed and actually do not have the disorder? What's the probability that they are prescribed lithium when the reality contradicts such a prescription? A reader unfamiliar to the field of psychiatric diagnostics may be appalled to find that, for many mental disorders, the probability of a patient not having a disorder upon receiving a positive diagnosis is not too different from getting the head side in a coin toss: a PPV of 50% is a norm in reality.

In the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project (Zimmerman, Ruggero, Chelminski & Young, 2008), the largest study into the overdiagnosis of BD, PPV was found to be a mere 43% from seven hundred psychiatric outpatients' diagnostic histories: Zimmerman et al. interviewed each patient using the Structured Clinical Interview for DSM Disorders (SCID) as the gold standard, and compared the diagnoses thus reached with clinicians' diagnoses of the same patient. They found that more than half the time, a patient who is diagnosed with BD in a community clinic turns out to not have the disorder. In a meta-analysis of 118 studies that analyzed the diagnostic accuracy of MDD by general practitioners, 19 of the studies were included for a robust evaluation of PPV, which averaged to be 42% (Mitchell, Vase, & Rao, 2009). Low PPV is common even for neurological disorders such as Alzheimer's disease (AD), which was found to be only 61% despite researchers raising the threshold for diagnosis (Lopez et al., 1999).

One major reason for the unsatisfactory PPV commonly seen in clinical diagnosis is the low prevalence of disorders in the population. To show the relationship between prevalence and predictive values more clearly, we can derive Equation 3 from dividing both the numerator and the denominator by *prevalence* in Equation 1, and derive Equation 4 from dividing both the numerator and the denominator in Equation 2 by *1-prevalence.*

$$\text{PPV}= \frac{sensitivity}{sensitivity+(1-specificity)*(\frac{1}{prevalence}-1)} \qquad \text{Equation 3.}$$

$$\text{NPV}= \frac{specificity}{(1-sensitivity)*\frac{prevalence}{1-prevalence}+specificity} \qquad \text{Equation 4.}$$

As can be seen in Equation 3. and 4., PPV and NPV are greatly influenced by prevalence: using a particular diagnostic instrument with certain sensitivity and specificity, the higher the prevalence of the disorder in question, the higher the PPV, the lower the NPV. Mental disorders such as BPD are relatively rare in the population: as mentioned above, the median prevalence is only 1.6%. Even for something as common as MDD, the point prevalence is estimated to be only 3% (Carney, 1988) : at any given time, 97% of the population would not have clinical depression. Clinicians can understand the relationship between prevalence and PPV intuitively—out of every 100 BPD diagnoses made in the general population, half may be false positive, simply because there are not that many people who have the disorder; however, 100 such diagnoses out of a group of people who experienced physical/sexual abuse in their childhood, within whom we know for a fact that BPD prevalence is high, there will be much fewer false positives—in other words, higher prevalence produces higher PPV, even when the same diagnostic criteria with the same sensitivity/specificity is applied.

Given the low prevalence of mental disorders in the population, what are some solutions to the low PPV problem? Researchers in the past have tried by revising the diagnostic criteria and raising the threshold for diagnosis to make it more difficult to warrant a diagnosis. Such is the strategy of the revising of DSM from DSM III (1980) to DSM 5 (2013). However, as shown above, this strategy is not working, and PPVs of diagnosing various disorders stay around 50%. Phelps & Ghaemi (2012) demonstrated the failure of this strategy by analyzing data from the largest study of the accuracy of BD diagnosis, which included more than 700 patients seen in community clinics. In that

study, sensitivity of diagnosing BD was 70%, specificity was 86%, PPV was 43% and NPV was 95%. Assuming the prevalence of BD to be 10% (higher than most estimates for psychiatric outpatient population), they held specificity constant, at the same level as the above data, and varied sensitivity, which is equivalent to tightening diagnostic criteria and lifting the threshold for diagnosis. As sensitivity increases, PPV only increases modestly; at the unrealistically high sensitivity of 100%, PPV barely reached 50% (Figure 2a). Manipulating specificity while keeping sensitivity at the 70% level has somewhat better results at increasing PPV: as can be seen in Figure 2b, PPV increments exponentially, and eventually reaches 100% with specificity at 100%. However, due to the restraint of prevalence rate at 10%, this exponential increment is only apparent after sensitivity reaches 90%. Consequently, even at the unattainably high level of sensitivity of 90%, PPV is only around 60%.

Phelps and Ghaemi's (2012) manipulation of sensitivity and specificity clearly demonstrates that past efforts at tightening diagnostic criteria are not effective at achieving high PPVs. They offered an alternative solution: increasing the prior probability in making diagnoses. The problem with existing DSM criteria is not that it doesn't have good sensitivity and specificity (at 87% and 70%), but that due to the restraint of low prevalence, they cannot reach a satisfactory PPV. This is where the Bayesian theorem helps save the day.

Chapter Three

Applying Bayesian Theorem to Statistical Diagnosis

Bayesian inference offers a way to improve our estimate of the probability of events, or in this context, the probability of a certain patient having a particular disorder. The Bayesian understanding of probability is that we as humans can hold prior beliefs of the probability of a certain event, and can update our estimate of the probability with the acquisition of pieces of new evidence. Clinicians may be already using Bayesian statistics intuitively. When a patient comes into a physician's office, presenting noticeable sadness and loss of interest in activities, the physician may already form a belief of the probability that this patient is suffering from MDD. As they interview the patient, they learn that the patient has also been experiencing insomnia and loss of appetite, and therefore may increase their estimate of the likelihood of MDD. As the patient then describes suffering from diminished ability to concentrate or make decisions in a timely fashion, the physician further increases their confidence of the patient having MDD. Such a process of constant updating our belief upon accumulating evidence is, in spirit, what we call Bayesian inference.

As elaborated in Chapter One, humans are not very accurate, nor able to perform with reasonable stability in this inference process. Apart from the reasons already enumerated, there is also the downfall [NOT SURE WHAT YOU ARE IMPLYING] of conditional probability and the difficulty to calculate them intuitively. In a probability space (Figure 3), the probability of a person having both a disorder (D) and a particular symptom (S), noted as $p(D \cap S)$, can be expressed in two ways (Equation 5. and Equation 6.), where $p(D|S)$ is the conditional probability of a patient having disorder D given the knowledge that they exhibit symptom S. In comparison, $p(S|D)$ is the conditional

probability that a patient manifests symptom S, given the knowledge that they truly have disorder D.

$p(D \cap S)=p(D|S)*p(S)$ Equation 5.

$p(D \cap S)=p(S|D)*P(D)$ Equation 6.

Linking the two equations together, we get:

$p(D|S)*p(S)= p(S|D)*P(D)$ Equation 7.

which leads to:

$p(D|S)= p(S|D)* p(D)/p(S)$ Equation 8.

Equation 8. is often referred to as Bayes Theorem, which gives foundation to the diagnostic method advocated in this thesis. If we have knowledge of the prevalence of a disorder in the general population, p(D), the prevalence of a particular symptom in the general population p(S), as well as the prevalence of the symptom in the disordered population p(S|D), we can arrive at the probability of one particular patient having the disorder D, given that they exhibit symptom S.

In reality, p(S) is not always readily available, due to limitations in epidemiological studies. However, we can derive from the epidemiology literature p(S|¬D), which is the prevalence of symptom S in the healthy population (i.e., people who do not suffer from disease D). From $p(S)=p(S|D)*p(D)+ p(S|\neg D)* p(\neg D)$, where p(¬D) is the probability of anyone not having disorder D, which equals to 1-p(D), we can derive Equation 9:

$$p(D|S) = \frac{p(S|D)* p(D)}{p(S|D)* p(D)+p(S|\neg D)* p(\neg D)}$$ Equation 9.

A valuable aspect to the application of Bayesian theorem in clinical diagnosis is that it enables us to iteratively update posterior probabilities given accumulating pieces of evidence (in the case of clinical diagnosis, accumulating pieces of information regarding patients' symptoms). Without any prior information on the symptoms patient A has, the probability of that person having disorder D is equal to the prevalence rate of D in the general population: p(D). If we know that patient A demonstrates symptom S_1, we can use Equation 7. to update this probability to $p(D|S_1)$: $p(D|S_1)=$ $\frac{\square(\square 1|\square)*\square(\square)}{\square(S1|\square)*\square(\square)+\square(\square 1|\neg\square)*\square(\neg\square)}$. If we also know that patient A demonstrates symptom S_2, we can substitute p(D) with $p(D|S_1)$ in the numerator of Equation 5 (our prior prevalence is now not based on the general population, but based on the population that have symptom S_1). Therefore, the probability that patient A has disorder D given that they demonstrate S_1 and S_2 is updated to be:

$p(D|S_1,S_2)=\frac{\square(\square 2|\square)*\square(\square|\square 1)}{\square(S2|\square)*\square(\square|\square 1)+\square(\square 2|\neg\square)*\square(\neg\square|\square 1)}$.

Repeating the use of Equation 5., as we gather more information on the series of symptoms that A suffers from $S=\{S_1, S_2, S_3 \ldots S_i \ldots S_n\}$, we can derive:

$$p(D|S)=\frac{p(D)*\prod_{i=1}^{n} p(Si|D)}{p(D)\prod_{i=1}^{n} p(Si|D)+p(\neg D)\prod_{i=1}^{n} p(Si|\neg D)]} \qquad \text{Equation 10.}$$

In theory, Equation 10. offers a way to accurately calculate the probability of the presence of a disorder given data on the prevalence rates of different symptoms in disordered vs healthy populations. How does this calculation fare in practice? Othmer et al. (2007) used three significant predictors to diagnose BD: onset of depression before age 25, family history of mania, and the presence of psychotic symptoms (Table 1.)

Using Equation 8 and the above data, we can calculate the predicted probability of an MDD patient also having BD when they have one of the above predictors, only S_1, when they have both S_1 and S_2 , as well as when they have all three predictors: S_1+S_1+S_1 (Table 2).

As Table 2. shows, the probabilities calculated using Bayesian Equations clearly delineate the trend that, as a patient demonstrates more and more relevant symptoms, they have a higher and higher likelihood of having a particular disorder. The Bayesian probabilities also conform to the actual data, as shown in column 3 of Table 2.

Chapter Four

Implementing the Expert System

So far, we have established two main problems that hinder the accuracy of differential diagnosis in current practice: 1) diagnoses made by clinicians, without the help of explicit statistical deduction with a clearly delineated scheme, are highly inaccurate. A corpus of literature has established that statistical diagnosis is superior to clinical diagnosis. 2) The current diagnostic criteria of mental disorders do not generate satisfactory PPVs. Tightening DSM criteria by raising the threshold for meeting diagnoses is not working to solve this problem, evidenced by the consistently low PPVs for different versions of DSM (see Pfohl, Coryell, Zimmerman, & Stangl, 1986; Phelps & Ghaemi, 2012). Bayesian inference circumvents these two problems by quantifying the

likelihood of a patient having specific mental disorders based on accumulating pieces of evidence.

The idea of using the Bayesian statistical method itself is nothing new. As early as 1986, Widiger, Frances, Warner, and Bluhm successfully used Bayesian inference to improve the diagnostic accuracy of BPD and Schizotypal Personality Disorder (SPD). Unfortunately, partly because Bayesian statistics are not familiar to clinicians, and the specific calculation of probabilities may be cumbersome, their study into the use of Bayesian methods in mental disorder diagnosis did not gain as much popularity as it deserves, and till this day, Bayesian methods have yet to be used in practice.

Many advances and trends in the field of healthcare and diagnostics since Widiger et al's study in 1986 have made the popularization of Bayesian methods a possibility today. Information systems such as Electronic Health Records (EHC) and the computerized clinical decision support system (CDSS) provide the environment for Bayesian diagnoses to thrive. The development of Bayesian algorithms based on CDSS goes as far back as the 1970s: deDombal and colleagues (1972) developed a CDSS that uses symptom prevalence rates as input to make diagnoses for acute abdominal pain, and correct diagnoses were reached in 91.8% patients (correct diagnoses were confirmed in subsequent abdominal surgical procedures), versus 65% to 80% correct diagnoses by clinicians. Despite such early interest in Bayesian CDSS, and the fact that such systems have been shown to significantly improve clinical practice two thirds of the time in a systematic review (Hunt, Haynes, Hanna, & Smith, 1998), the widespread adoption of CDSS is a recent and growing phenomenon. There are many reasons for the slow

adoption of CDSS in particular, and medical information systems in general, in hospitals and other clinical settings.

Historically, before computers were put to wide use in hospitals, CDSS could not be popularized. More pertinent to today's world where computers are ubiquitous, challenges come from compatibility and convenience issues. Physicians tend to be reluctant to adopt CDSS if it adds extra steps to their routine clinical practice. Incompatibility issues may require physicians to manually enter medical information that already exists in another record system. In the fast-paced hospital environment, these additional steps dampen physicians' enthusiasm to switch to CDSS. Many recent studies have identified key aspects that make a CDSS successful for popularization (e.g. Kawamoto, Houlihan, Balas, & Lobach, 2005), and the teams that develop CDSS are paying more attention to elements that determine a viable CDSS, such as integration into a clinician's workflow, providing support at the time of decision making, and providing treatment recommendations in addition to assessments.

The demonstrated benefits of computer-based medical information systems, as well as the medical research community's recognition of the importance of evidence-based care, have prompted policy makers to encourage the utilization of health information technology. The Obama administration passed the Health Information Technology for Economic and Clinical Health Act (HITECH) in 2009, which financially incentivizes medical providers to implement medical information system technology. In the past decade, especially after HITECH, most hospitals in the US have moved from a traditional paper based system to electronic health records: 96.7% of non-federal US

acute care hospitals in the US are using federally certified EHR systems in 2014 (see *Figure 4*).

It is against such a background of a more computerized healthcare system, as well as the increasing popularity of CDSS, that this paper discusses the development of a computerized expert system that integrates Bayesian statistics into the diagnosis of mental disorders. The expert system will present to the clinician all symptoms of target disorders, and the clinician simply checks the symptoms presented in the patient; the system will then automatically calculate the probability of the patient having each mental disorder, on which clinicians can then base their treatment strategies. The system will also be compatible with EHR systems, so that relevant medical records could be directly integrated into Bayesian calculations.

The first step, a minimally viable product (MVP) of such an expert system has been developed with Matlab. (*Appendix A* shows the source code). It presents a graphic user interface (GUI) which offers a step-wise guide for clinicians to check the patient's presenting symptoms. Prevalence rates of each symptom in the disordered population and in the general population are collected from peer-reviewed literature from databases such as PubMed and PsychInfo, and Bayesian probability equations (e.g. Equation 10.) are used to calculate the likelihood of the patient having BD type I, BD type II, BPD, ADHD, and AsPD, respectively. The minimally viable expert system has been tested with some hypothetical cases: *Appendix B* shows screenshots of each step of the diagnostic process using this system for one hypothetical patient. The probability of BD and BPD are calculated based on prevalence rates from epidemiological studies cited in *Appendix C*.

Due to a paucity of symptom-specific prevalence data for ADHD and AsPD, however, this current expert system used DSM criteria to reach a binary diagnosis for these two disorders.

Chapter Five

Future Steps and Caveats

Before the expert system can be scaled to include all DSM disorders and used in practice, the accuracy of the system needs to be tested against current clinical diagnosis methods. Studies testing diagnostic accuracy will follow the common practice of using SCID as the gold standard, as a remedy to the "ground truth" problem. As the output of a Bayesian expert system will be probabilities, rather than a binary diagnosis of presence versus absence of disorders, how clinicians should make treatment decisions based on these probabilities needs further study. Although an intuitive approach is to set a threshold probability for positive versus negative diagnosis, that may not leverage all the advantages of the Bayesian method. The possibility of setting thresholds to treatment actions, instead of thresholds for binary diagnoses should be investigated: a patient who has 50% likelihood of having BD probably requires a different prescription strategy than someone with 90% likelihood of having BD. The thresholds will also vary by disorder, as the treatment for varying disorders entail different tradeoffs. Does a 55% probability of having BD type II justify prescribing lithium? What about 65%? 70? Does a 50% probability of having BPD justify the expense of DBT? Such questions require extensive research.

Precisely what treatment strategy should be recommended at what likelihood levels is a direction for further research.

There are many advantages to the Bayesian expert system. Apart from the improved accuracy inherent in its use of statistical rather than clinical methods, having likelihood as output will help clinicians devise more individualized, and likely more effective treatment plans. The system requires minimal input from clinicians (they only have to click a few buttons on the screen), and can be seamlessly integrated in their encounter with patients. The expert system could be deployed on different digital platforms, such as desktop and tablet computers, and could be integrated with widely used EHR systems such as EPIC. Information from patients' health records such as family history could be automatically input into the system when making diagnoses.

There are, of course, factors that will potentially hinder the popularity of such an expert system. Bayesian methods are not familiar to the clinical community. Psychologists who have received some training are usually taught frequentist statistics, and may not be open to the Bayesian way of thinking, and therefore unwilling to switch to the Bayesian expert system. This is a major reason that a Bayesian framework has not been widely accepted yet. In the development of MVP, I also noticed the scarcity of literature on prevalence rates of specific symptoms in diseased versus general/healthy populations. This completes a vicious cycle where the lack of epidemiological data impedes the shift to a Bayesian framework, and the reluctance to accept such a shift makes research less interested in investing in detailed epidemiological studies. It will take

a paradigm shift for Bayesian inference to be eventually accepted and used by clinicians in practice.

Although the development of the Bayesian expert system is motivated by previous evidence on unsatisfactory clinical diagnoses, arguments can be made against its use. First, the argument could be made that the standards for testing diagnostic accuracy used in previous research favor statistical diagnosis. As a remedy to the "ground truth" problem, the gold standard used in studies of mental disorder diagnosis has been the Structured Clinical Interview for DSM Disorders (Lobbestael, Leurgans, & Arntz, 2011), which is statistical in nature (if you possess a certain number of criteria, you are pronounced to have a certain disorder). This may bias the accuracy of results towards statistical models. However, there has to be some standard to measure different diagnostic methods against, and negating the superiority of statistical diagnosis is denying the value of quantitative research in epidemiology. Even in treatment efficacy studies, we have to have some kind of standard to measure the improvement after treatment, and this is implemented with psychometric scales, which are statistical in nature themselves. If we neglect what the numbers and statistics tell us, we'll only be following the opposite path of evidence-based medical practice.

There is also the argument against the current categorization of mental disorders. High comorbidity and the profuse shared symptoms between different mental disorders cast doubt on the current categorization. In many cases, various disorders respond to the same type of treatment: for example, methylphenidate, a primary pharmacological treatment for ADHD, has been shown to effectively alleviate BPD symptoms (Golubchik,

Sever, Zalsman, & Weizman, 2008), while the Dialectical Behavioral Therapy and Cognitive Behavioral Therapy for PTSD and BPD have many shared components including exposure-based treatment (Becker & Zayfert, 2001). It could be argued that instead of putting efforts into the differential diagnoses of these disorders based on the current categorization system, resources should be pooled into the study of their shared symptom etiology, and the potential development of a dimensional model of mental disorders, where disorders are not understood as distinct categories, but manifestations of patients' unique standing on several key dimensions such as impulsivity, mood lability, and attention sustainability. Yet the move towards a dimensional model will likely take a long time, given that current etiology research has only offered us a glimpse into the complex construct of mental disorders: one semi-dimensional diagnostic model is included in the current version of Diagnostic and Statistical Manual of Mental Disorders (DSM-5) as an alternative to the current categorization of personality disorders, however, to the author's knowledge, since the publication of DSM-5, there has been no study that investigates the performance of this alternative model in clinical settings, let alone the impact of this alternative model on effective treatment. In previous versions of DSM (DSM I in 1952 and DSM II in 1968), psychiatrists' then understanding of disorder etiology strongly influenced their categorization. The transition to the symptom driven categorization starting from DSM III actually improved the performance of the diagnosing symptom in generating less comorbidity and better reliability (Tyrer, 2014). It's possible that with recent advances in more scientific understanding of mental disorders based on a biopsychosocial model, we may eventually succeed with

reattempting the etiology based diagnostic system. That being said, in the time for such a model to surface, improvements in diagnostic accuracy of the current categorization system can generate more immediate benefits to psychiatric patients.

There are many limitations to the minimally viable expert system developed with this paper. First, the Bayesian algorithm used in this program assumes the symptoms to be conditionally independent, which is rarely true for clinical disorders. For example, a person with impulsive traits in general would also be more likely to exhibit para-suicidal behaviors, both of which are symptoms of BPD. Because we have not come close to a comprehensive understanding of the links between symptoms for various mental disorders, assuming statistical independence is a useful simplification to make the problem of diagnosis solvable. Of course, more advanced Bayesian methods such as the Markov random field, which does not assume independent probabilities, may be used in more complex expert systems that could even achieve higher diagnostic accuracy.

Another disadvantage of the current software is that it only allows yes or no (check/uncheck) options for each symptom. Because of this, many statistics from literature that had ratings for symptoms instead of straight cut-offs were not used as eligible data sources. In the future development of the expert system, a ratings based approach should also be considered.

The prevalence data used in the current expert system regards the general population, rather than the population seen in primary care settings. As symptom prevalence rates are usually much higher in primary care settings, the probabilities output

by the current system may be an underestimate. The reason that general population prevalence rates were used, again, goes back to the paucity of data in the literature.

For distinction between BD type I and type II, the current system stuck to the DSM random cutoff of mania duration that determines mania vs hypomania. This is because the prevalence rates seen in literature often do not make a distinction between BD I and BD II. All these statistics that are lacking in the literature clearly demonstrates the necessity for more research effort in epidemiology studies. This project and the promise of Bayesian diagnosis may serve as a rallying cry to encourage more epidemiological research into symptom-specific prevalence rates.

Reference

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C: American Psychiatric Association.

Basco, MR, Bostic, JQ, Davies, D, et al.: Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 157:1599-1605, 2000

Becker, C. B., & Zayfert, C. (2001). Integrating DBT-based techniques and concepts to facilitate exposure treatment for PTSD. *Cognitive and Behavioral Practice, 8*(2), 107-122.

Carney, R. M., Rich, M. W., Freedland, K. E., Saini, J., Simeone, C., & Clark, K. (1988). Major depressive disorder predicts cardiac events in patients with coronary artery disease. Psychosomatic Medicine, 50(6), 627-633.

Chilakamarri, J. K., Filkowski, M. M., & Ghaemi, S. N. (2011). Misdiagnosis of bipolar disorder in children and adolescents: A comparison with ADHD and major depressive disorder. *Ann Clin Psychiatry, 23*(1), 25-29.

Coulston, C. M., Tanious, M., Mulder, R. T., Porter, R. J., & Malhi, G. S. (2012). Bordering on bipolar: the overlap between borderline personality and bipolarity. *Australian and New Zealand Journal of Psychiatry, 46*(6), 506-521.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

De Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. Br Med J, 2(5804), 9-13.

DeShong, H. L., & Kurtz, J. E. (2013). Four factors of impulsivity differentiate antisocial and borderline personality disorders. *Journal of personality disorders, 27*(2), 144.

Eyler, A. E. (2007). Diagnostic Bias in Men's Healthcare. *Clinical Men's Health: Evidence in Practice*, 13.

Fennig S, Naisberg-Fennig S, Craig TJ, et al. Comparison of clinical and research diagnoses of substance use disorders in a first-admission psychotic sample. Am J Addiction 5: (1) 40-48 Winter 1996

Ferrer, M., Andión, Ó., Matalí, J., Valero, S., Navarro, J. A., Ramos-Quiroga, J. A., . . . Casas, M. (2010). Comorbid attention-deficit/hyperactivity disorder in borderline patients defines an impulsive subtype of borderline personality disorder. *Journal of personality disorders, 24*(6), 812.

Francois, D., Roth, S. D., & Klingman, D. (2015). The Efficacy of Pharmacotherapy for Borderline Personality Disorder: A Review of the Available Randomized Controlled Trials. *Psychiatric Annals, 45*(8), 431-437.

First, Michael B., Spitzer, Robert L, Gibbon Miriam, and Williams, Janet B.W.: Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P) New York: Biometrics Research, New York State Psychiatric Institute, November 2002.

Fish, V. (2004). Some gender biases in diagnosing traumatized women. *Bias in psychiatric diagnosis*, 213-220.

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.

Golubchik, P., Sever, J., Zalsman, G., & Weizman, A. (2008). Methylphenidate in the treatment of female adolescents with co-occurrence of attention deficit/hyperactivity disorder and borderline personality disorder: a preliminary open-label trial. *International clinical psychopharmacology, 23*(4), 228-231.

Gross, R., Olfson, M., Gameroff, M., Shea, S., Feder, A., Fuentes, M., ... & Weissman, M. M. (2002). Borderline personality disorder in primary care. Archives of Internal Medicine, 162(1), 53-60.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs: General and Applied, 79*(9), 1.

Grove, W. M. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of clinical psychology, 61*(10), 1233-1243.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment, 12*(1), 19.

Heffernan, K., & Cloitre, M. (2000). A comparison of posttraumatic stress disorder with and without borderline personality disorder among women with a history of childhood sexual abuse: etiological and clinical characteristics. *The Journal of nervous and mental disease, 188*(9), 589-595.

Hirschfeld, R., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry*.

Hunt, D. L., Haynes, R. B., Hanna, S. E., & Smith, K. (1998). Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. Jama, 280(15), 1339-1346.

Jørgensen, C. R., Freund, C., Bøye, R., Jordet, H., Andersen, D., & Kjølbye, M. (2013). Outcome of mentalization□based and supportive psychotherapy in patients with borderline personality disorder: a randomized trial. *Acta Psychiatrica Scandinavica, 127*(4), 305-317.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 330(7494), 765.

Linehan, M. M., Comtois, K. A., Murray, A. M., Brown, M. Z., Gallop, R. J., Heard, H. L., . . . Lindenboim, N. (2006). Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of general psychiatry, 63*(7), 757-766.

Lopez, O. L., Litvan, I., Catt, K. E., Stowe, R., Klunk, W., Kaufer, D. I., ... & DeKosky, S. T. (1999). Accuracy of four clinical diagnostic criteria for the diagnosis of neurodegenerative dementias. Neurology, 53(6), 1292-1292.

Leichsenring, F., Leibing, E., Kruse, J., New, A. S., & Leweke, F. (2011). Borderline personality disorder. *The Lancet, 377*(9759), 74-84.

Lobbestael, J., Leurgans, M., & Arntz, A. (2011). Inter□rater reliability of the Structured Clinical Interview for DSM□IV Axis I disorders (SCID I) and Axis II disorders (SCID II). *Clinical psychology & psychotherapy, 18*(1), 75-79.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet, 374(9690), 609-619.

Othmer, E., DeSouza, C. M., Penick, E. C., Nickel, E. J., Hunter, E. E., Othmer, S. C., ... & Hall, S. B. (2007). Indicators of mania in depressed outpatients: a retrospective analysis of data from the Kansas 1500 study. *Journal of Clinical Psychiatry, 68*(1), 47-51.

Phelps, J., & Ghaemi, S. (2012). The mistaken claim of bipolar 'overdiagnosis': solving the false positives problem for DSM□5/ICD□11. *Acta Psychiatrica Scandinavica, 126*(6), 395-401.

Pfohl, B., Coryell, W., Zimmerman, M., & Stangl, D. (1986). DSM-III personality disorders: Diagnostic overlap and internal consistency of individual DSM-III criteria. Comprehensive Psychiatry, 27(1), 21-34.

Ruggero, C. J., Zimmerman, M., Chelminski, I., & Young, D. (2010). Borderline personality disorder and the misdiagnosis of bipolar disorder. *Journal of psychiatric research, 44*(6), 405-408.

Salzbrenner, S., & Conaway, E. (2009). Misdiagnosed Bipolar Disorder Reveals Itself to be Posttraumatic Stress Disorder with Comorbid Pseudotumor Cerebri: A Case Report. *Psychiatry (Edgmont), 6*(8), 29.

Shea, M. T., Zlotnick, C., & Weisberg, R. B. (1999). Commonality and specificity of personality disorder profiles in subjects with trauma histories. *Journal of personality disorders, 13*(3), 199.

Singh, M. K., DelBello, M. P., Kowatch, R. A., & Strakowski, S. M. (2006). Co□ occurrence of bipolar and attention□deficit hyperactivity disorders in children. *Bipolar disorders, 8*(6), 710-720.

Spitzer, R. L. (1983). Psychiatric diagnosis: are clinicians still necessary?.*Comprehensive psychiatry*, *24*(5), 399-411.

Swartz, M., Blazer, D., George, L., & Winfield, I. (1990). Estimating the prevalence of borderline personality disorder in the community. *Journal of personality disorders, 4*(3), 257.

Tyrer, P. (2014). A comparison of DSM and ICD classifications of mental disorder. *Advances in psychiatric treatment, 20*(4), 280-285.

Widiger, T. A., Frances, A., Warner, L., & Bluhm, C. (1986). Diagnostic criteria for the borderline and schizotypal personality disorders. Journal of Abnormal Psychology, 95(1), 43.

Wingo, A. P., & Ghaemi, S. N. (2007). Frequency of stimulant treatment and of stimulant-associated mania/hypomania in bipolar disorder patients. *Psychopharmacology bulletin, 41*(4), 37-47.

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine* (Vol. 569): John Wiley & Sons.

Zlotnick, C., Johnson, D. M., Yen, S., Battle, C. L., Sanislow, C. A., Skodol, A. E., . . . Bender, D. S. (2003). Clinical features and impairment in women with borderline personality disorder (BPD) with posttraumatic stress disorder (PTSD), BPD without PTSD, and other personality disorders with PTSD. *The Journal of nervous and mental disease, 191*(11), 706-713.

Zanarini, M. C., Frankenburg, F. R., Dubo, E. D., Sickel, A. E., Trikha, A., Levin, A., & Reynolds, V. (1998). Axis I comorbidity of borderline personality disorder. *American Journal of Psychiatry*.

Zlotnick, C., Johnson, D. M., Yen, S., Battle, C. L., Sanislow, C. A., Skodol, A. E., . . . Bender, D. S. (2003). Clinical features and impairment in women with borderline personality disorder (BPD) with posttraumatic stress disorder (PTSD), BPD without PTSD, and other personality disorders with PTSD. *The Journal of nervous and mental disease, 191*(11), 706-713.

Zimmerman, M., Ruggero, C. J., Chelminski, I., & Young, D. (2008). Is bipolar disorder overdiagnosed?. Journal of Clinical Psychiatry, 69(6), 935-940.

Table 1. Symptom prevalence in MDD+BD patients versus MDD patients. Data source: Othmer et al. (2007)

	MDD patients with BD N=200	MDD patients without BD N=544
S_1: onset of depression before age 25	74%	55%
S_2: family history of mania	34.5%	15.5%
S_3: presence of psychotic symptoms	40%	15.3%

Table 2. Calculated probability of BD with accumulating symptoms, compared to actual percentage of BD patients who have 0, 1, 2, 3 symptoms.

	Bayesian predicted probability of having BD	Actual percentage of patients with BD
0 symptom	14.7%	14.7%
S_1	18.9%	19.3%
S_1+S_2	35.5%	48.8%
S_1+S_2+ S_3	75%	67%

Note. The actual percentages are from Othmer et al. (2007), based on diagnoses by SCID as the golden standard .

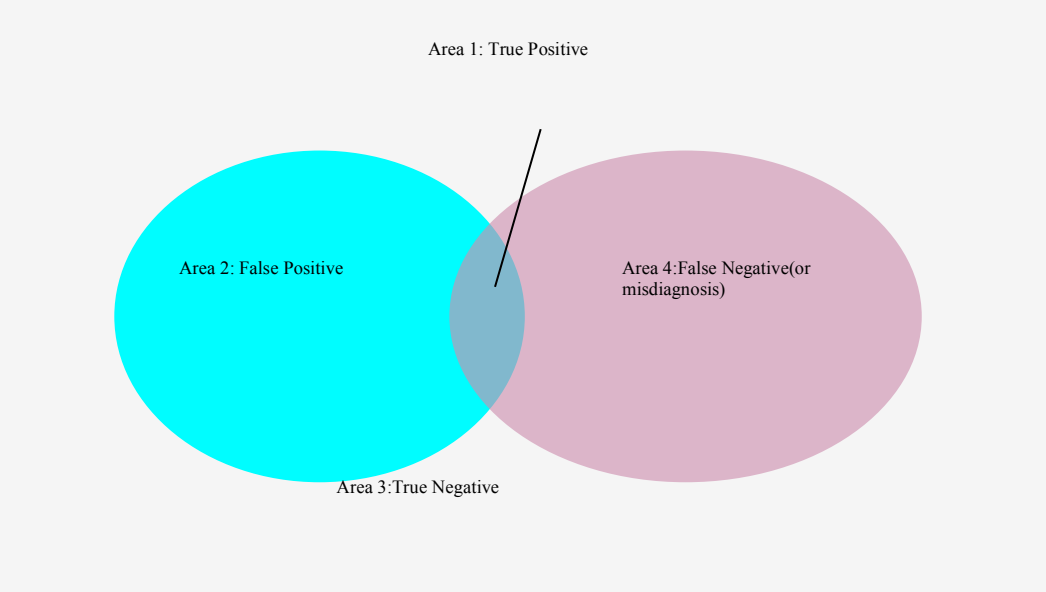


Figure 1. Venn Diagram showing four combinations of diagnosis vs actual condition.

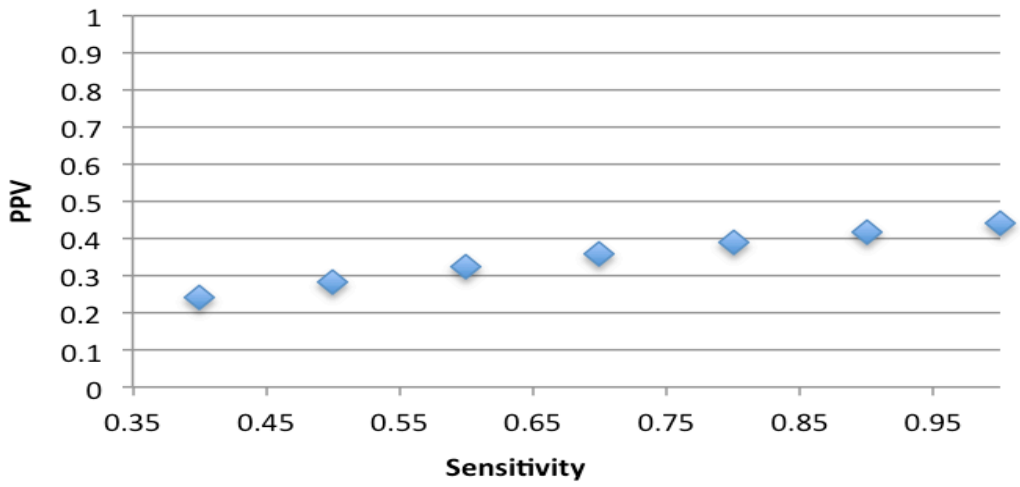


Figure 2a. Reproduction of Phelps and Gaemi's data manipulation, showing the relationship betwen PPV and sensitivity. Specificitiy is set at 86%, and prevalence at 10%.

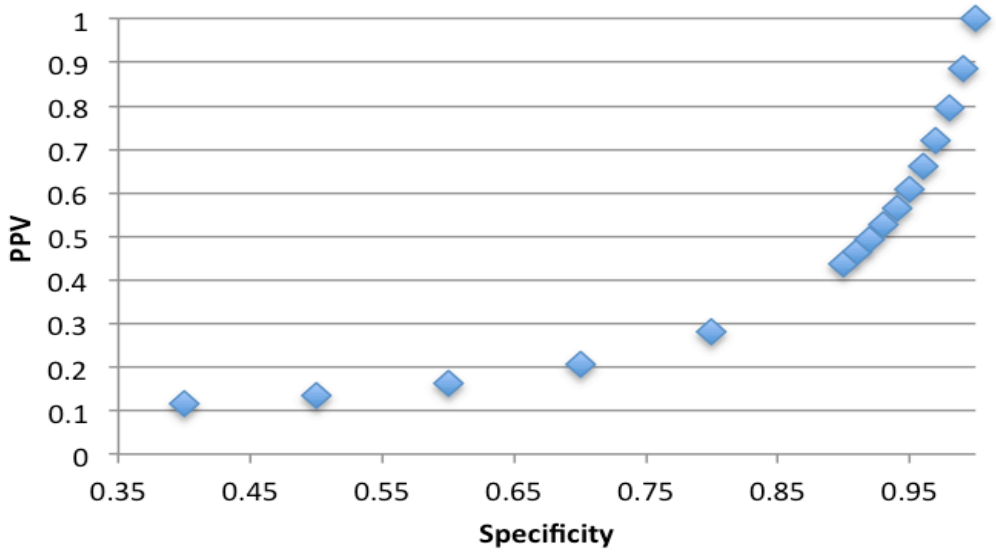


Figure 2b. Reproduction of Phelps and Gaemi's data manipulation, showing the relationship betwen PPV and specificity. sensitivity is set at 70%, and prevalence at 10%.

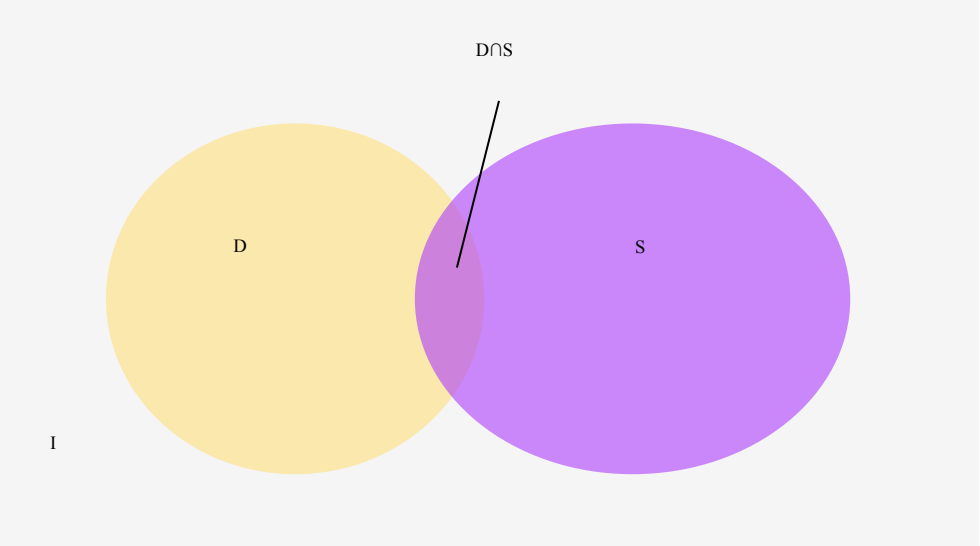


Figure 3. Disorder and One of its Symptom in a probability Space

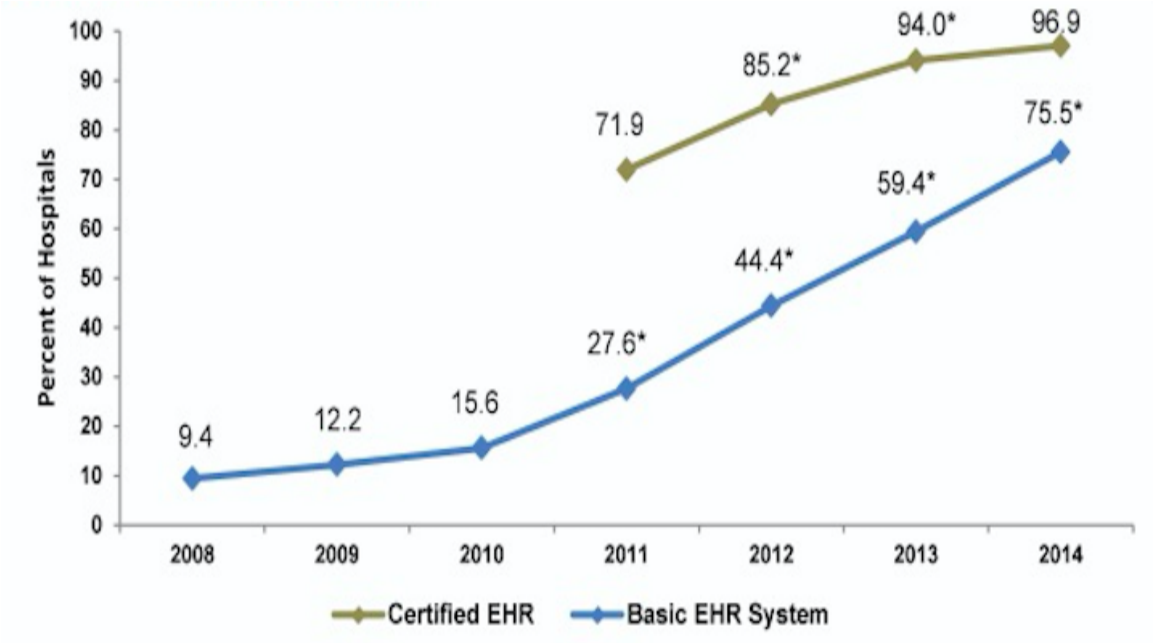
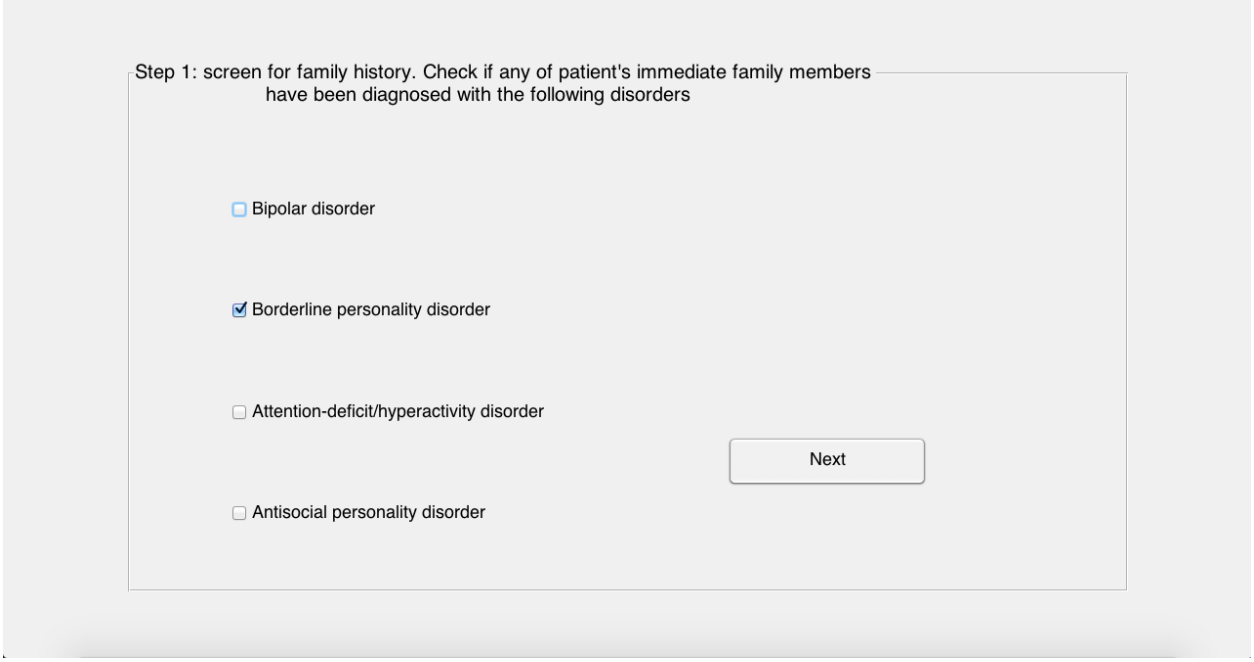
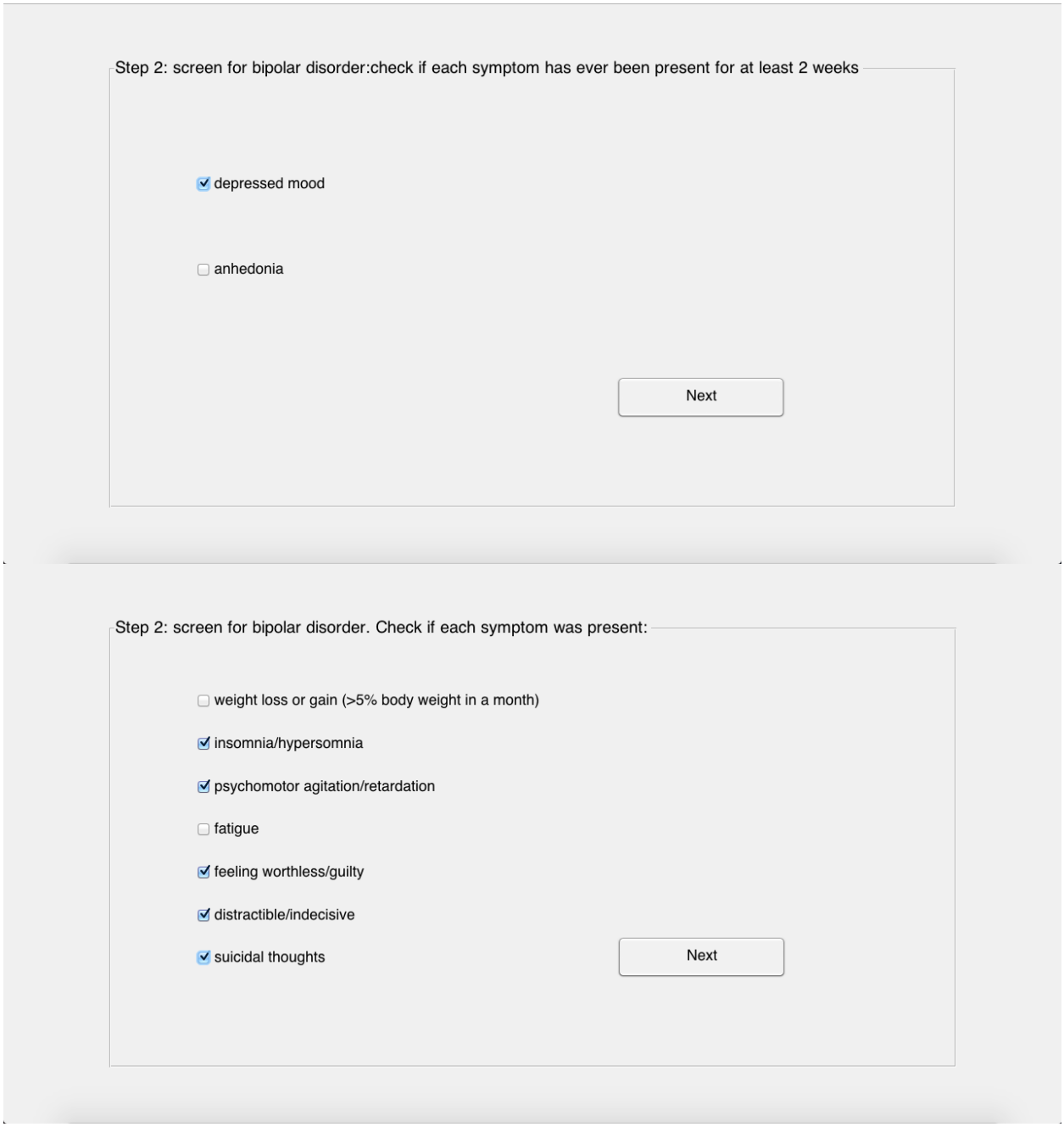


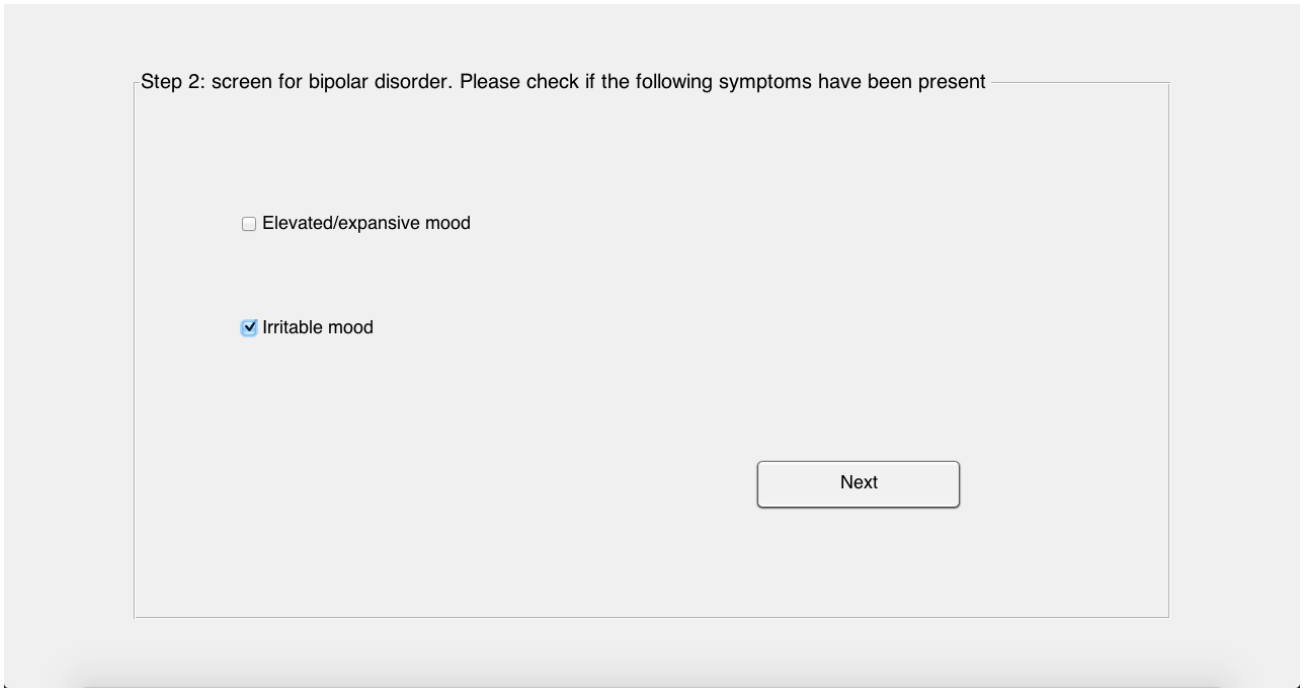
Figure 4. Percentage of non-federal acute care hospitals that adopted either basic EHR system or federally certified EHR system between 2008 and 2014. Source: The Office of the National Coordinator for Health IT (ONC, 2016). Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014.

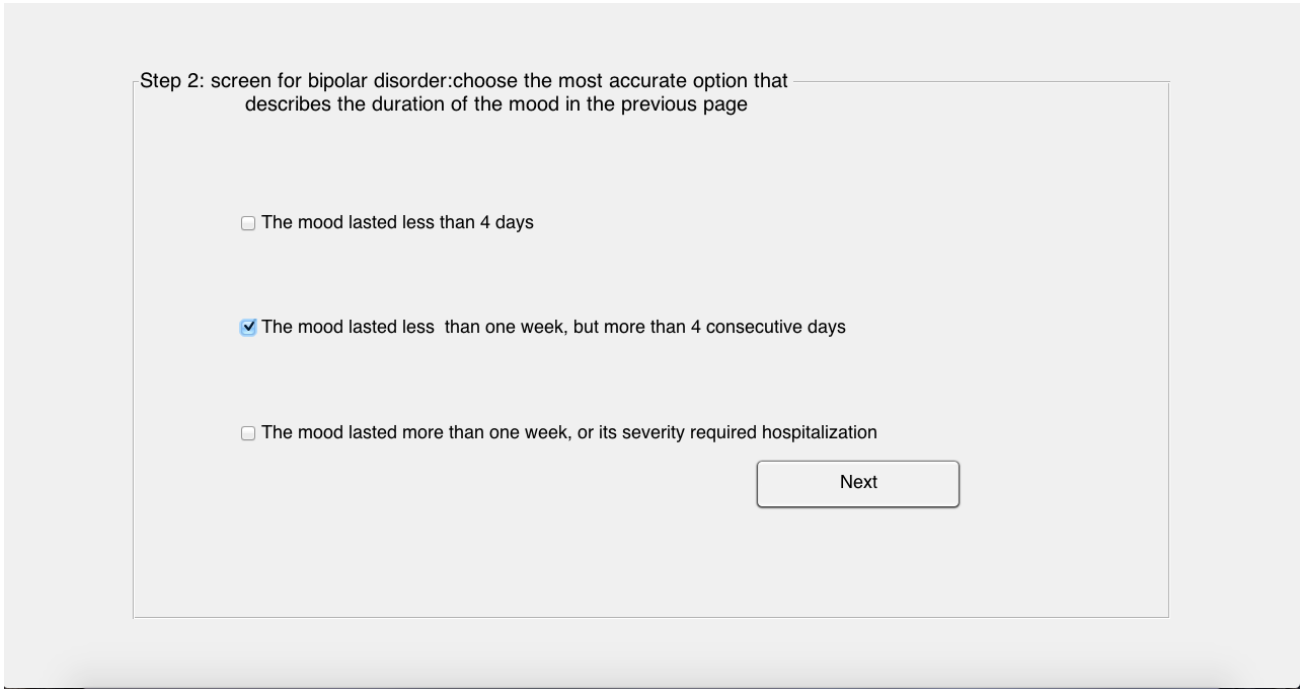
Appendix B

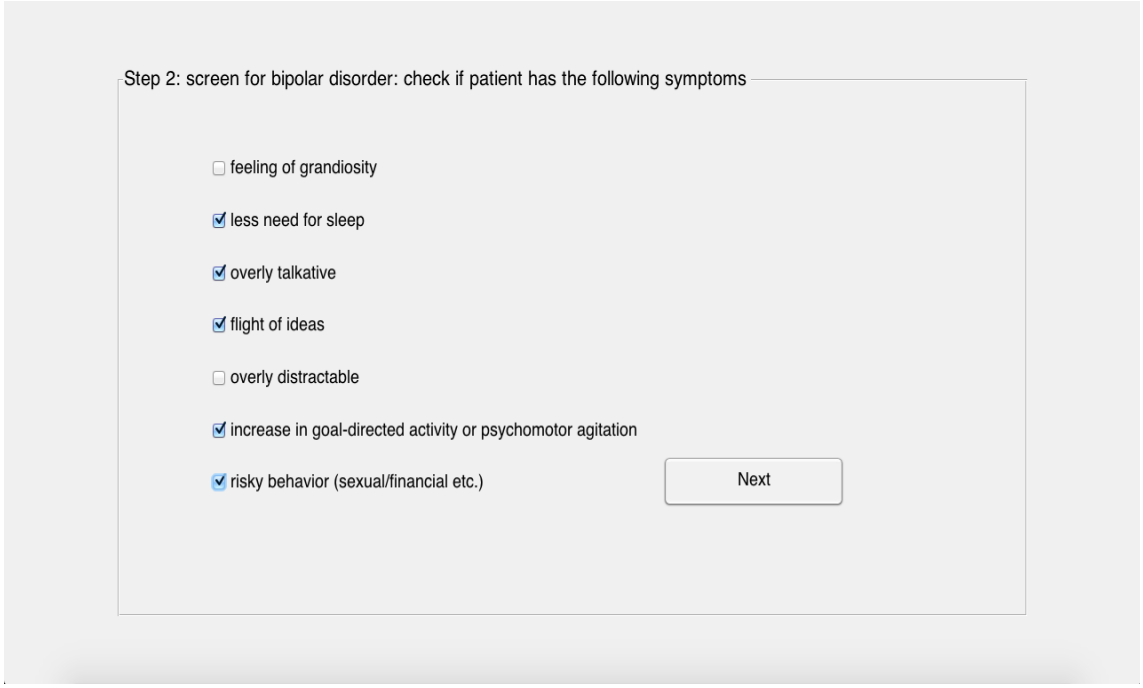
Screenshots of Diagnostic Process and Results

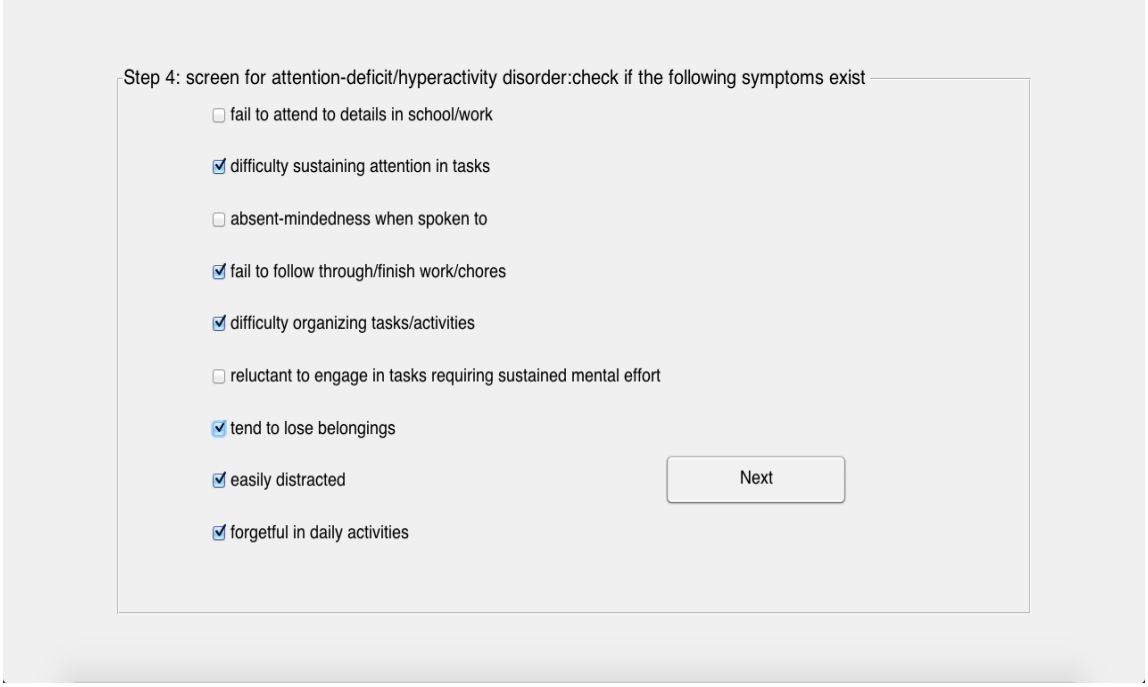


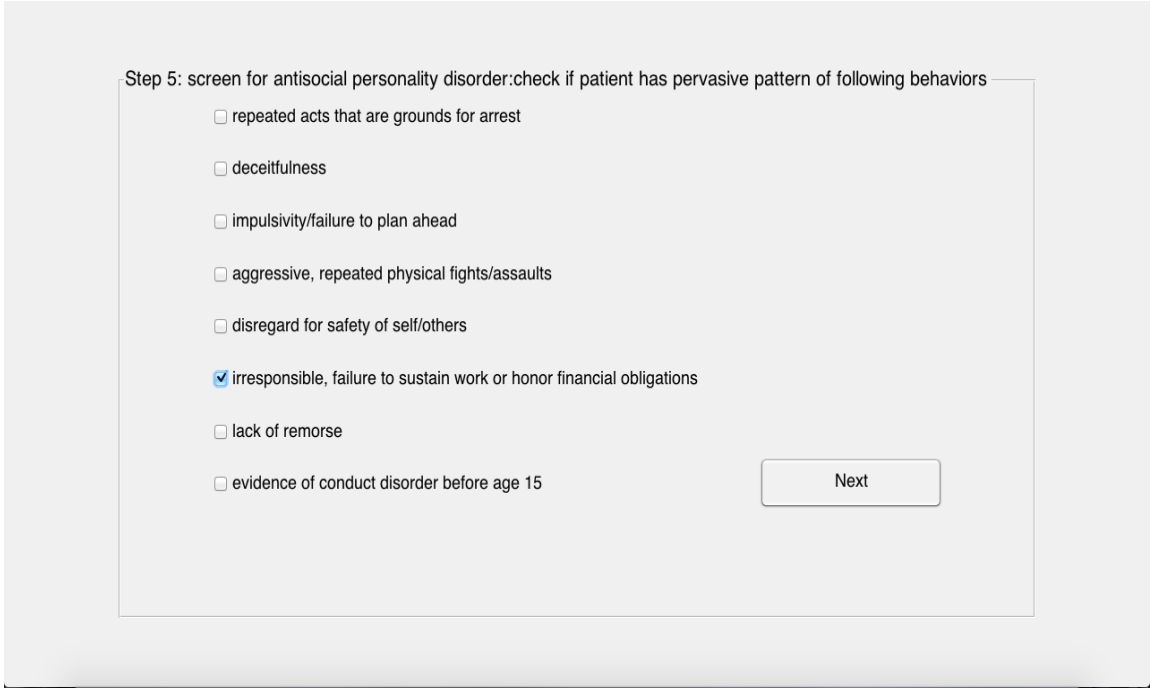

Step 2: screen for bipolar disorder:check if each symptom has ever been present for at least 2 weeks
depressed mood
anhedonia
Next
Step 2: screen for bipolar disorder. Check if each symptom was present:
weight loss or gain (>5% body weight in a month)
insomnia/hypersomnia
psychomotor agitation/retardation
fatigue
feeling worthless/guilty
distractible/indecisive
suicidal thoughts
Next

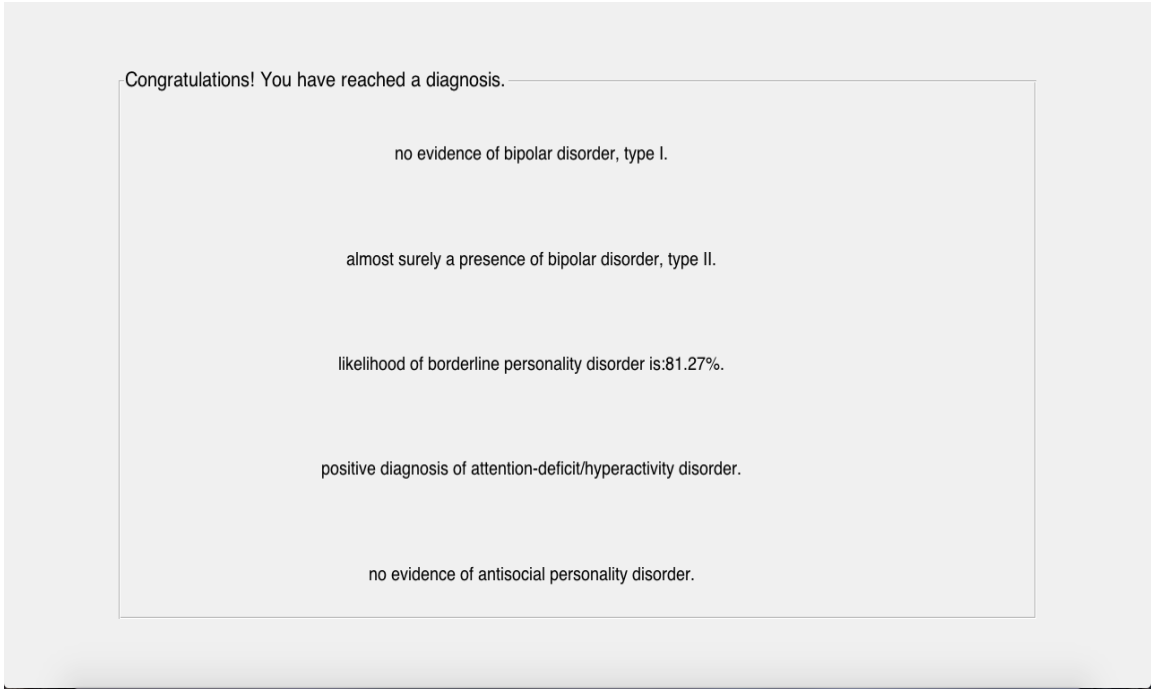

Step 2: screen for bipolar disorder. Please check if the following symptoms have been present
Elevated/expansive mood
Irritable mood
Next


Step 2: screen for bipolar disorder:choose the most accurate option that describes the duration of the mood in the previous page
The mood lasted less than 4 days
The mood lasted less than one week, but more than 4 consecutive days
The mood lasted more than one week, or its severity required hospitalization
Next


Step 2: screen for bipolar disorder: check if patient has the following symptoms
feeling of grandiosity
less need for sleep
overly talkative
flight of ideas
overly distractable
increase in goal-directed activity or psychomotor agitation
risky behavior (sexual/financial etc.)
Next


Step 4: screen for attention-deficit/hyperactivity disorder:check if the following symptoms exist
fail to attend to details in school/work
difficulty sustaining attention in tasks
absent-mindedness when spoken to
fail to follow through/finish work/chores
difficulty organizing tasks/activities
reluctant to engage in tasks requiring sustained mental effort
tend to lose belongings
easily distracted
forgetful in daily activities
Next


Step 5: screen for antisocial personality disorder:check if patient has pervasive pattern of following behaviors
repeated acts that are grounds for arrest
deceitfulness
impulsivity/failure to plan ahead
aggressive, repeated physical fights/assaults
disregard for safety of self/others
irresponsible, failure to sustain work or honor financial obligations
lack of remorse
evidence of conduct disorder before age 15
Next


Congratulations! You have reached a diagnosis.
no evidence of bipolar disorder, type I.
almost surely a presence of bipolar disorder, type II.
likelihood of borderline personality disorder is:81.27%.
positive diagnosis of attention-deficit/hyperactivity disorder.
no evidence of antisocial personality disorder.

Appendix C

Symptom Specific Prevalence Rates used in the MVP

BD

Depression: population prevalence: 17%

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence. *Journal of affective disorders*, *29*(2-3), 85-96.

BD prevalence: 100%

feelings of grandeur: 5

population prevalence 12%

Morey, L. C. (1985). A comparative validation of the Foulds and Bedford hierarchy of psychiatric symptomatology. *The British Journal of Psychiatry*,*146*(4), 424-428.

BD prevalence 59%

Knowles, R., McCarthy-Jones, S., & Rowse, G. (2011). Grandiose delusions: a review and theoretical integration of cognitive and affective perspectives.*Clinical psychology review*, *31*(4), 684-696.

reduced need for sleep

population prevalence 16.6%

Breslau, N., Roth, T., Rosenthal, L., & Andreski, P. (1996). Sleep disturbance and psychiatric disorders: a longitudinal epidemiological study of young adults. *Biological psychiatry*, *39*(6), 411-418.

BD prevalence 84%

Harvey, A. G., Talbot, L. S., & Gershon, A. (2009). Sleep disturbance in bipolar disorder across the lifespan. *Clinical Psychology: Science and Practice*, *16*(2), 256-277.

racing thoughts/pressure of speech

Population prevalence: 30%

Goodwin FK, Jamison KR: Manic Depressive Illness. New York: Oxford University Press; 1990.

BD: 72%

Goodwin FK, Jamison KR: Manic Depressive Illness. New York: Oxford University Press; 1990.

Distractibility/Psychomotor Agitation

Population prevalence: 1.5%

bipolar: 90%

Balázs, J., Benazzi, F., Rihmer, Z., Rihmer, A., Akiskal, K. K., & Akiskal, H. S. (2006). The close link between suicide attempts and mixed (bipolar) depression: implications for suicide prevention. *Journal of affective disorders*,*91*(2), 133-138.

Impulsivity (impulse control disorder)

Population prevalence: 3%

BD: 13%

McElroy, S. L., Pope, H. G., Keck, P. E., Hudson, J. I., Phillips, K. A., & Strakowski, S. M. (1996). Are impulse-control disorders related to bipolar disorder?. *Comprehensive psychiatry*, *37*(4), 229-240.

Depression

population: 6.5%

Kandel, D. B., & Davies, M. (1982). Epidemiology of depressive mood in adolescents: an empirical study. *Archives of general psychiatry*, *39*(10), 1205.

BD 50.3%

Borderline Personality disorder

paranoia/dissociation

Population prevalence: 33%

Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Fowler, D., ... & Dunn, G. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *The British Journal of Psychiatry*,*186*(5), 427-435.

BPD 68%

McGlashan, T. H., Grilo, C. M., Sanislow, C. A., Ralevski, E., Morey, L. C., Gunderson, J. G., ... & Stout, R. L. (2005). Two-year prevalence and stability of individual DSM-IV criteria for schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders: toward a hybrid model of axis II disorders. *American Journal of Psychiatry*.

Impulsivity

Population prevalence: 39%

BPD 89%

Unstable, intense relations

Population prevalence: 39%

BPD: 77%

Inappropriate anger population

45%

BPD

87%

Identity disturbance bpd3

population

13%

BPD

49%

Affective instability bpd6

population

49%

BPD

79%

Intolerance of being alone bpd1

population

39%

BPD

70%

Physically self-damaging acts bpd5

population

36%

BPD

91%

Emptiness and boredom bpd7

population

16%

BPD

58%

Widiger, T. A., Frances, A., Warner, L., & Bluhm, C. (1986). Diagnostic criteria for the borderline and schizotypal personality disorders. *Journal of Abnormal Psychology*, *95*(1), 43.

ASPD (this disorder is especially tricky, as there are almost no studies investigate the prevalence of individual criterion either in the disordered or general population.)

acts that are grounds for arrest

general population

10%

ASPD

35.3%

Black, D. W., Gunter, T., Loveless, P., Allen, J., & Sieleni, B. (2010). Antisocial personality disorder in incarcerated offenders: Psychiatric comorbidity and quality of life. *Ann Clin Psychiatry*, *22*(2), 113-20.

substance abuse

general

10

Black, D. W., Gunter, T., Loveless, P., Allen, J., & Sieleni, B. (2010). Antisocial personality disorder in incarcerated offenders: Psychiatric comorbidity and quality of life. *Ann Clin Psychiatry*, *22*(2), 113-20.

Family History

BD 2% vs 7.5%

BPD 1.6% vs 9.1%

Bandelow, B., Krause, J., Wedekind, D., Broocks, A., Hajak, G., & Rüther, E. (2005). Early traumatic life events, parental attitudes, family history, and birth risk factors in patients with borderline personality disorder and healthy controls. *Psychiatry Research*, *134*(2), 169-179.