Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not

Eugenie Jackson¹ Christina Mendoza¹

¹Equivant/Northpointe Inc., United States of America

Published on: Mar 31, 2020

DOI: <https://doi.org/10.1162/99608f92.1b3dadaa>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

1. Introduction

[“The Age of Secrecy and Unfairness in Recidivism Prediction”](#) (Rudin, Wang, & Coker, 2020) treats a wide variety of topics. It includes a clever approach to determining the age dependency of a COMPAS Core Risk and Needs Assessment (hereafter Core RNA). Regrettably, we find that several mistakes mar this effort. We will not be able to discuss all of our disagreements or engage all of the benefits of the Core RNA. Due to restrictions of time and space, we address in detail only what we consider to be the most pressing of these problems.

First, we seek to rectify the authors’ misconceptions about what the Core RNA is and how it is intended to be used. We provide brief biographies of its primary developers, along with a description of others who have supported the COMPAS tools over the years. Next, we describe the independent validation studies that have been conducted. These studies have consistently shown the efficacy and fairness of the Core RNA using established methods. We then address the topic of transparency, noting that Northpointe, Inc., is engaged in efforts to copyright its models to enhance their transparency. We also discuss data security and privacy. Finally, we describe several significant deficiencies we find in the authors’ approach.

2. Risk and Needs Assessments: Background

As the authors note, the Core RNA is “a product of years of painstaking theoretical and empirical sociological study” (Rudin et al., 2020). We begin by discussing some of the theoretical bases of RNAs.

A foundation of modern criminal justice practice is the risk-need-responsivity (RNR) principle. The RNR principle holds that programming to reduce recidivism is most beneficial to those at highest risk, when it is targeted to address verified needs, and administered in a manner well-suited to an offender’s learning style (Andrews, Bonta, & Hoge, 1990). Indeed, some studies have found that administering programming to low-risk individuals may be associated with increased recidivism rates (Andrews & Bonta, 2016). Following this principle, the complete Core RNA consists of two risk scales, the General Recidivism Risk Scale (GRRS) for recidivism in general and the Violent Recidivism Risk Scale (VRRS) for violent recidivism, as well as a set of need scales that can be configured to meet an agency’s requirements to enable the RNR principles. Designed to work together, the risk and need scales are used in individual case management to inform supervision and programming decisions.

Risk factors are constructs that cannot typically be measured by a single item, but require multiple items. Most of the Core risk and need scales contain four items or more. Despite the authors’ claim that more items entail more error, having several items in a well-developed scale is actually more accurate than a single item when measuring a construct (Nunnally & Bernstein, 1994). Northpointe tests its scales for internal consistency, reliability, and construct validity. Detailed information about the validity of the Core RNA scales is available in Section 3.2 of the Practitioner’s Guide to COMPAS Core (Northpointe, 2019).

Two types of risk factors, static and dynamic, may be included in creating a recidivism risk scale. Static factors, such as past Criminal History, tend to be unchanging and cannot respond to treatment. Dynamic risk factors, also known as criminogenic needs, are changeable and responsive to treatment. Dynamic risk factors are further classified as stable or acute. Stable dynamic factors, like Criminal Personality, can change, but only slowly, while acute dynamic factors, such as mood, may change quickly (Hanson & Harris, 2000).

Different criminological approaches advocate for varying combinations of factor types in risk assessment. For example, the state of California uses its own risk assessment tool, the California Static Risk Assessment (CSRA), based on the Washington State Model (Barnoski & Aos, 2003). It is composed entirely of static risk factors. Other assessments, such as the Level of Service Inventory-Revised (LSI-R) and the Ohio Risk Assessment System (ORAS) use a combination of static and dynamic risk factors. The Core RNA also uses static and dynamic risk factors, with the important distinction that the use of dynamic factors is restricted to only those that are stable. This restriction results in risk scores that may change slowly over time. See Desmarais & Singh (2013) for a more thorough discussion of risk factors and some of the risk assessments used throughout the criminal justice system.

2.1. The Core Risk and Need Assessment: Its Structure and Uses

There are many misconceptions expressed in the article about what the Core RNA is and how it is intended to be used. In this subsection, we provide a map that shows how the Core RNA fits into the larger set of tools provided in the Northpointe Suite software, while describing its structure and uses.

The Northpointe Suite is a decision-support software package designed to offer criminal justice agencies a wide range of case management tools, including tools to assess risk and need. The Northpointe Suite contains many widely used RNAs such as the LSI-R (Andrews & Bonta, 1995), as well as specialty assessments like the Texas Christian University Drug Screen (Simpson, Joe, Knight, Rowan-Szal, & Gray, 2012) and the Static-99/R (Phenix et al., 2017). The Northpointe Suite also contains internally developed tools, including a risk scale for pretrial release, a jail classification tool, a recidivism risk screen, typology tools to assist in internal classification, an acute risk factor assessment for case management, agency-specific RNAs, and the Core and Reentry RNAs. Each tool is designed to be used for individuals at a specific decision point that may have different information requirements across the criminal justice system. Contained within the complete Core RNA are the GRRS, the VRRS, and numerous need scales (e.g., Substance Abuse; Residential Instability) that agencies use as appropriate to help ensure optimal supervision levels and allocation of resources.

GRRS and VRRS: Like all recidivism risk assessments, the purpose of the GRRS and VRRS is to predict future reoffending given information about an individual at the time of assessment. Their purpose is not to describe or diagnose someone as being a recidivist or nonrecidivist, because recidivism is not a preexisting condition (Royal Statistical Society Section on Statistics and the Law, 2018).

The authors mistakenly state that “(t)o compute the COMPAS raw scores, Northpointe collects 137 variables from a questionnaire, computes a variety of subscales, and finally linearly combines the subscales and two age variables—the defendant’s age at the time of the current offense and age at the time of the first offense—to compute the raw risk scores.” There is no reason offered for the assumption that all 137 questions are used in the VRRS and GRRS. Indeed, the authors overlook the evidence in their own Table 1. Table 1 asserts that the “Total” number of features for the subscales referenced by their equation for VRRS in section 2.1 is $5 + 9 + 12 = 26$, of which the authors have $3 + 8 + 0 = 11$. A similar critique holds for their repeated claim that the GRRS involves 137 questions.

Other tools for understanding risk: A theme of our present discussion is that recidivism is a complex subject and it is to everyone’s advantage to obtain a nuanced and complete recidivism risk profile. For example, the VRRS does not contain items related to current violence because the data have not supported that current violence helps to predict violence in general. Thus, it is recommended that the Current Violence scale be used in conjunction with the GRRS and VRRS to inform case management (Northpointe, 2019).

Criminologists have long been aware of the possibility of predicting risk with a small number of variables. Northpointe’s recidivism risk screen (RRS) is a five-item risk screen that may be given if there is an institutional requirement for a shorter risk assessment procedure. The items that make up this screen are age, age at first arrest, number of prior arrests, employment status, and prior parole revocations, (Northpointe, 2019). The short recidivism risk screen may be used by an agency as a triage tool to assess whether someone is low risk and, thus, may not need to complete a full Core RNA. This approach saves time and resources.

Core need scales: The Core need scales describe facets of an individual that are relevant for correctional practice. Needs that are noncriminogenic, such as housing, may also be included in an agency’s configured set of need scales. These are included because it has been shown that treating these needs can support the treatment of criminogenic needs:

[C]riminal justice professionals are likely to have a very difficult time addressing a participant’s antisocial attitudes or delinquent peer interactions if he or she is living on the street, suffering from a severe mental illness, or experiencing acute withdrawal symptoms from drugs or alcohol. Ignoring or delaying attention to these concerns is not a realistic course of action if one wishes to reduce crime and rehabilitate criminal justice-involved persons. (Marlowe, 2018)

The majority of items in the full Core RNA were explicitly included to accurately screen an individual for areas of greatest need. According to the RNR principle, without this information, agencies are not able to put their resources to best use nor to avoid the possibility that individuals may receive inappropriate, even harmful, treatment (Lowenkamp & Latessa, 2004).

Appropriate use: The Core RNA scale set was developed for use with individuals who have been convicted and are beginning a probation, parole, or prison sentence. During development, recidivism for the GRRS was measured as any new arrest within 3 years of probation or parole start; while in contrast, recidivism for the VRRS was measured as any new person offense arrest within 3 years of probation or parole start.

A widely accepted position is that courtroom decisions should not be based solely on a recidivism risk score. The guidelines issued by the National Center for State Courts (NCSC) regarding how risk assessments are to be used in evidence-based sentencing are clear on this (Casey, Warren, & Elek, 2011). NCSC's first guideline states: "Risk and need assessment information should be used in the sentencing decision to inform public safety considerations related to offender risk reduction and management. It should not be used as an aggravating or mitigating factor in determining the severity of an offender's sanction" (Casey et al., 2011, p. 11). That is, a person's risk scores should not have any effect on the sentence imposed. The second guideline states: "Risk and needs assessment information is one factor to consider in determining whether an offender can be supervised safely and effectively in the community" (Casey et al., 2011, p. 14). Other specific factors must be considered when determining "whether an offender is a good candidate for community supervision" (Casey et al., 2011, p. 14).

2.2. The Developers of the Core RNA

The original version of the Core RNA was developed during the mid-1990s by Tim Brennan and Dave Wells, cofounders of Northpointe, Inc. Dr. Brennan is a psychometrician whose research includes groundbreaking work on runaways, women's pathways to crime, and offender classification. He has held positions at the University of Colorado-Boulder in the Institute of Cognitive Science and at Georgia State University in the Department of Criminal Justice and Criminology. Wells's background in corrections led to his interest in criminal justice research. Brennan and Wells started Northpointe because they believed there were better ways to get offenders the help they needed while streamlining management tools for overburdened agencies.

Over the years, the Northpointe research team has incorporated the expertise of academics in the fields of criminology, sociology, social work, psychology, computer science, and statistics. Additionally, former DOC practitioners, wardens, state agency administrators, and other dedicated professionals from the field of criminal justice have supported Northpointe's work. Northpointe's internally developed tools are trusted in the agencies where they are used thanks to the expertise and integrity of these criminal justice professionals.

3. Testing and Validation

When any scale or assessment is added to or changed in the Northpointe Suite, the Research and Development departments work together to test relevant combinations of inputs for the instrument and the corresponding scores. Missing items are also generated by well-established imputation methods and the scores are carefully analyzed. This work complements the extensive risk-assessment quality-assurance practices of criminal justice agencies.

A key background factor is that Northpointe’s models have been validated by multiple independent researchers as well as by the Northpointe research department with consistently positive results. Table 1 lists some of the validation results obtained by researchers for both Core and Reentry versions of the GRRS and VRRS in different geographical locations and using different definitions of recidivism. For example, the column labeled “Person” refers to assaultive offenses following release during fixed study times. The studies by Farabee, Zhang, Roberts, and Yang (2010), Flores, Lowenkamp, and Bechtel (2016), Lansing (2012), Reich, Picard-Fritsche, Barber Rioja, and Rotter (2016), and Rong and Matthews (2018) are of particular note since the authors are independent of Northpointe. Area under the curve (AUC) scores are reported for each study. “The consensus in the field of recidivism research seems to be that AUC values below 0.65 are poor, 0.65 to 0.69 are fair, 0.70 to 0.75 are good, and 0.76 and above are excellent” (Northpointe, 2019). AUCs indicate the discriminative ability of the scales. Even more relevant in evaluating a recidivism risk scale is its predictive ability (Levy, 2018; Royal Statistical Society Section on Statistics and the Law, 2018). For example, in Rong and Matthews (2018), the authors report a 13.3% 3-year recidivism rate in the Core low risk category, a 23.9% 3-year recidivism rate in the medium risk category, and a 43.3% 3-year recidivism rate in the high-risk category, when analyzing the relationship between GRRS levels and recidivism. More information about the positive predictive values of the risk scales are included in the referenced articles.

Table 1. Summary of AUC results for the general recidivism risk scale (GRRS) and violent recidivism risk scale (VRRS) in several outcomes studies.

Study	N	Year	Any Arrest	Felony	Person	Supervision Failure
NY Probation ¹	2,328	2009	0.680	0.700	0.710	
NY Probation ²	13,993	2012	0.710			
MDOC Reentry ³	25,347	2011		0.710	0.700	0.690
MDOC Probation ⁴	21,101	2011		0.670	0.740	0.710
CDCR Reentry ⁵	25,009	2010	0.700		0.650	
Broward Jail ⁶	6,172	2016	0.710		0.710	
Mental Health Court ⁷	242	2016	0.730			

Santa Barbara Probation ⁸	5,363	2017	0.722		0.672	0.702
Riverside Probation ⁹	4,435	2018	0.694		0.636	0.692
Massachusetts DOC ¹⁰	1,813	2018	0.691			

4. Transparency

Striking a balance between protecting the investments made in developing the risk assessments and allowing increased transparency has been a goal of Northpointe for some time. Northpointe and its parent company, equivant, are pursuing copyrights for the GRRS and VRRS. A feature that has been widely ignored is that every agency using these instruments already has full access to all risk variables, logic, scoring processes, and guidelines for the appropriate uses of risk assessment procedures. Agencies are expected to use the risk assessments in the manner they were designed to be used—in the appropriate population, at the right point in the criminal justice process, and within the constraints of the legal requirements of their particular jurisdictions. Hence, quite appropriately, ultimate discretion about disclosure of scores to assessed individuals, or even whether they will be used at all, lies with duly constituted governing authorities.

4.1. Not a Black Box but a Statistical Model With Theoretically Justifiable Features

The developers and researchers who support the GRRS and VRRS work to understand the complexity of criminological theory and the data that inform these theories. The blend of stable dynamic and static risk factors to generate reliable risk scores is based in criminological theory and data (Andrews, Bonta, & Wormith, 2006). For example, as the authors note, the VRRS is a weighted combination of age, age at first arrest, Violence History, Noncompliance History, and Vocational/Educational need. Transformations familiar to every statistician are employed to improve the models. The low signal-to-noise ratio inherent in recidivism data further supports the use of statistical methods rather than machine learning (Harrell, 2019). Northpointe's approach simultaneously affirms the complexity of predicting recidivism while providing case managers the tools to understand the resulting scores.

The authors generally conflate complexity of method, complexity of model, interpretability, and secrecy. It might be argued, along lines suggested by Tollenaar and van der Heijden (2019), that a branch and bound algorithm like the one used in the CORELS method (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017) is not especially simple or interpretable. The authors have assumed that the Core RNA uses complex methods and/or models, is hard to interpret, and is opaque to users. None of this is correct. The Core RNA was

developed using widely established methods yielding simple models that are directly interpretable and fully transparent to the agencies that use it.

4.2. Data: Ownership, Security, Privacy

All risk and need data belong to the agencies who collect it. Many agencies have their own research and IT departments, and much of their time is spent monitoring data quality and ensuring confidentiality. Unlike credit agencies who may own clients' data, Northpointe owns no data and has no “control over criminal risk scores” (Rudin et al., 2020). This is a fiction the authors propagate. Northpointe complies with all agency requirements to ensure the security and privacy of the data it works with. Background checks and data-sharing agreements are usually required before work can begin on any project.

The Broward County assessment data used by the authors present several data-quality issues. The data were collected during 2013 and 2014 (Angwin, Larson, Mattu, & Kirchner, 2016). The assessment records were matched with criminal history records in 2016. These data were never fully described in any ProPublica news story. There is no way to verify the quality of these data. For example, it is not known how accurately race and ethnicity were coded in the assessment records. The methods used to compose the sample are hidden. It is possible that selection effects biased the sample. Thus, the representativeness and overall quality of the sample is questionable. Moreover, the records within the Broward County data set are predominantly pretrial cases, despite the fact that the Core RNA is intended for use on a postconviction population. Finally, no one in the data set received the full Core RNA; only the Core GRRS and VRRS were administered.

Northpointe does not publish the names of individuals from the data it uses and objects to the 2016 posting of names on GitHub from the data used in Angwin et al. (2016).

5. Criticism

We have many criticisms of the article. We address the most important ones here. We begin with some general comments. We then describe some misleading assertions by the authors. Finally, we discuss a series of significant deficiencies that we relate to the concept of fairness.

To begin with, we expect the authors to be above using sensationalist tactics like suggesting that the Core RNA has anything to do with crimes like the murder of the man in Westervelt (2017).¹¹ A similar tactic is employed when the authors write, “In the past, there have been documented cases where individuals have received incorrect COMPAS scores based on incorrect criminal history data (Wexler, 2017a, b) and have had no mechanism to correct it after a decision was made based on that incorrect score.” The use of the plural “cases” and two citations belie the fact that there is a single individual being described. Moreover, the item discussed in those stories is not part of either risk assessment, but part of a need scale. Another sensationalist tactic is the authors' repeated claim that the General and Violent risk scales have 137 items when their own table shows that together they only have 40.

Among the deficiencies in the authors' arguments, we note first a surprising assumption regarding the age dependence in risk scores. The authors have taken a clearly informal description of the VRRS score in the *Practitioner's Guide to COMPAS Core* (Northpointe, 2019) for a complete technical description of the VRRS model. This guide is written for practitioners and is not intended to be a technical document. Discussions of appropriate variable transformations are beyond its scope and would not further its goals; however, we note that the skewed age variable is an ideal candidate for a normalizing transformation (see Figure A3 in authors' Appendix)¹². Regardless, the GRRS model is not the same as the VRRS model and there is no reason to assume that no other age-dependent variables enter the GRRS model. This assumption is vital to their insinuation that a supposed opacity has allowed Northpointe's documentation to misstate the risk scores' functional dependence on age for years with no checks.

Second, we note that the authors assume the GRRS score has no additional negative terms and no additional age dependence over the violence risk score. Both assumptions are critical to their analysis of their attempted reconstruction of the GRRS score. Although the authors qualify them as relatively weak, there is no justification offered for these assumptions.

A third particularly significant error hampers their effort to validate their attempted isolation of age dependence in the risk scores. In Section 2.2 they compare the performance of several models both with and without age inputs, but always with the "age at first arrest" input. As the skewed age histogram in their appendix indicates, age at first arrest is a good proxy for age because most of the population is young. Including the age at first arrest input confounds their model comparison because both sets of compared models effectively see age as an input. The authors rely on this flawed finding to suggest the GRRS score does not rely on inputs like "rate of crimes committed over time" (Rudin et al., 2020).

We note next a pair of erroneous assumptions the authors appear to make. It appears the authors conflate number of arrests with number of charges. These are two very different kinds of things. A single arrest can be associated with many charges. Additionally, the authors appear to conflate long criminal history with extensive criminal involvement. Again, these are two different kinds of things, as can be seen by considering a case of high age with early first arrest yet low criminal involvement. In Section 3, the authors present low-risk scores that are associated with high-charge tallies (or "counts") in the context of discussing putative misattribution of high criminal involvement as low risk. Nowhere in this discussion is there any direct evidence of high criminal involvement presented. Apparently, counts are being mistaken as evidence of high criminal involvement, which in turn is mislabeled as a long criminal history.

We come next to the highly significant assumptions that the GRRS and VRRS must be complex, difficult to interpret, and entirely opaque. None of these assumptions is warranted. As we have stressed earlier, the GRRS and VRRS are fully interpretable and transparent to the agencies that use them. In Brennan et al. (2009), a publicly available peer-reviewed article in a relevant criminal justice journal, the risk variables, statistical

methods (logistic regression and survival analyses) used for the Core RNA, and validation results are explicitly described. Similar descriptions are provided in training sessions to all of Northpointe’s users.

Finally, the authors make several assumptions about recidivism risk without any stated empirical argument whatsoever. These seemingly bare intuitions include that a long criminal history should imply a high risk and that current arrest should contribute to it. In Section 3, the authors even suggest that the VRRS “seems” wrong because a boosted tree approach disagrees, apparently with no validation beyond hold-out tactics on the Broward County data and for which accuracy even in this limited testing is not discussed. There have been multiple published suggestions from those outside the field of criminology that a ‘simple’ feature set like age, number of prior arrests, perhaps sex, should suffice for an adequate recidivism prediction. The authors make a similar suggestion in the supplied citations for their claim that any number of off-the-shelf methods can dispense with “bespoke” methods founded in criminological expertise. Conversely, the GRRS and VRRS recidivism models are informed by decades of empirical assessment; this assessment incorporates external validation for transportability using the clinically appropriate measure of positive predictive value rather than retrospective discrimination alone; the GRRS and VRRS have been described and illustrated at national criminal justice and corrections conferences over this span; and their predictive validity and theoretical foundations have been peer-reviewed in relevant professional journals, and confirmed by a growing number of external research teams. The authors’ seeming reliance on mere opinion over the collective effort of an academic discipline engaged in empirical research is jarring.

This brings us to the connection we announced between these numerous deficiencies and the notion of procedural fairness. We suggest that care and due diligence are critical ingredients to the procedure of academic research and publishing, not least because that is what fairness to readers requires.

6. Conclusion

We have provided a primer on what the Core RNA is and how it is intended to be used. We linked this description to the instruments’ creators, supporters, and users. We have also described how these instruments have been validated in studies, demonstrating their efficacy and fairness using established methods. We have addressed the topics of the transparency and interpretability of the Core Risk and Need Assessment. We also discussed data security and privacy. Finally, we described several significant deficiencies we found in the authors’ approach.

We wish to reiterate that recidivism is a complex phenomenon, not one that can be solved by simply giving a person a single risk score nor by attempting to label a person a recidivist or nonrecidivist. Our efforts go toward supporting theoretically and empirically justified tools that improve outcomes for all members of society.

Acknowledgements

Thanks to Tim Brennan and Bill Dieterich of Northpointe, Inc., for their helpful suggestions.

Disclosure Statement

Eugenie Jackson and Christina Mendoza have no financial or non-financial disclosures to share for this article.

References

Andrews, D. A., & Bonta, J. (1995). *LSI-R: The level of service inventory—revised*. Toronto, Canada: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2016). *The psychology of criminal conduct* (6th ed.). London: Routledge.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17(1), 19–52. <https://doi.org/10.1177/0093854890017001004>

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, 52(1), 7–27. <https://doi.org/10.1177/0011128705281756>

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18, 8753–8830. <https://doi.org/10.1145/3097983.3098047>

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Barnoski, R., & Aos, S. (2003). *Washington's offender accountability act: An analysis of the Department of Corrections' risk assessment (document no. 03-12-1202)*. Olympia: Washington State Institute for Public Policy. Retrieved from [http://www.wsipp.wa.gov/ReportFile/847/Wsipp Washington-s-Offender-Accountability-Act-An-Analysis-of-the-Department-of-Corrections-Risk-Assessment Full-Report.pdf](http://www.wsipp.wa.gov/ReportFile/847/Wsipp%20Washington-s-Offender-Accountability-Act-An-Analysis-of-the-Department-of-Corrections-Risk-Assessment%20Full-Report.pdf)

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21-40. <https://doi.org/10.1177/0093854808326545>

Casey, P., Warren, R., & Elek, J. (2011). *Using offender risk and needs assessment information at sentencing: Guidance for courts from a national working group*. Williamsburg, VA: National Center for State Courts. Retrieved from https://www.ncsc.org/_data/assets/pdf_file/0019/25174/rna-guide-final.pdf

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., & Misra, S. (2018). The Public Safety Assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in Kentucky. <http://doi.org/10.2139/ssrn.3168452>

Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Lexington, KY: Council of State Governments Justice Center. Retrieved from <https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf>

Dieterich, W., Brennan, T., & Oliver, W. L. (2011). *Predictive validity of the COMPAS Core risk scales: A probation outcomes study conducted for the Michigan Department of Corrections*. Traverse City, MI: Northpointe.

Dieterich, W., Mendoza, C. M., & Brennan, T. (2017). *Discrimination and calibration of the Core VFO and NonVFO Risk Scales: Checking the accuracy of the risk models in updated probation validation samples*. Traverse City, MI: Northpointe.

Dieterich, W., Mendoza, C., Hubbard, D., Ferro, J., & Brennan, T. (2018). *COMPAS risk scales validation study: An outcomes study conducted for the Riverside County Probation Department*. Traverse City, MI: Northpointe.

Dieterich, W., Oliver, W., & Brennan, T. (2011). *Predictive validity of the reentry COMPAS risk scales: An outcomes study with extended follow-up conducted for the Michigan Department of Corrections*. Traverse City, MI: Northpointe.

Farabee, D., Zhang, S., Roberts, R. E., & Yang, J. (2010). *COMPAS validation study: Final report*. UCLA Integrated Substance Abuse Programs.

Flores, A. W., Lowenkamp, C. T., & Bechtel, K. (2016, July). *False positives, false negatives, and false analyses: A rejoinder*. Retrieved from https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf

Hanson, R. K., & Harris, A. J. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior*, 27(1), 6–35. <https://doi.org/10.1177/0093854800027001002>

Harrell, F. (2019, September 15). Road map for choosing between statistical modeling and machine learning. *Statistical Thinking*. Retrieved from <https://www.fharrell.com/post/stat-ml/>

Lansing, S. (2012). *New York State COMPAS-Probation Risk and Needs Assessment Study: Evaluating predictive accuracy*. Albany, NY: New York State Division of Criminal Justice Services, Office of Justice Research and Performance. Retrieved from https://www.criminaljustice.ny.gov/crimnet/ojsa/opca/compas_probation_report_2012.pdf

Levy, D. (2018). In machine learning predictions for health care the confusion matrix is a matrix of confusion. *Statistical Thinking*. Retrieved from <http://www.fharrell.com/post/mlconfusion/>

Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections, 2004*, 3–8. Retrieved from https://www.uc.edu/content/dam/uc/ccjr/docs/articles/ticc04_final_complete.pdf

Marlowe, D. B. (2018). *The most carefully studied, yet least understood, terms in the criminal justice lexicon: Risk, Need, and Responsivity*. Retrieved February 26, 2019, from <https://www.prainc.com/risk-need-responsivity/>.

Northpointe. (2019). *Practitioner's guide to COMPAS Core*. Traverse City, MI: Author. Retrieved from <http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf>

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Phenix, A., Fernandez, Y., Harris, A. J., Helmus, M., Hanson, R. K., & Thornton, D. (2017). *Static-99R coding rules, revised-2016*. Public Safety Canada (Sécurité Publique Canada). Retrieved from http://www.static99.org/pdfdocs/Coding_manual_2016_v2.pdf

Reich, W. A., Picard-Fritsche, S., Barber Rioja, V., & Rotter, M. (2016). *Evidence-based risk assessment in a mental health court: A validation study of the COM-PAS risk assessment*. New York, NY: Center for Court Innovation. Retrieved from <http://www.courtinnovation.org/sites/default/files/documents/COMPAS%20Validation%20Study.pdf>

Rong, J., & Matthews, H. (2018). *General Recidivism Risk Score levels: Three-year review analysis*. Massachusetts Department of Correction, Office of Strategic Planning and Research. Retrieved from https://www.mass.gov/files/documents/2018/05/03/Risk_Score_Analysis_2018.pdf

Royal Statistical Society Section on Statistics and the Law. (2018, November 8). Algorithms in the justice system: Some statistical issues. Retrieved from https://www.rss.org.uk/Images/PDF/influencing-change/2018/RSSsubmission_Algorithms_in_the_justice_system_Nov_2018.pdf

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.6ed64b30>

Simpson, D. D., Joe, G. W., Knight, K., Rowan-Szal, G. A., & Gray, J. S. (2012). Texas Christian University (TCU) short forms for assessing client needs and functioning in addiction treatment. *Journal of Offender Rehabilitation*, 51(1–2), 34–56. <https://doi.org/10.1080/10509674.2012.633024>

Tollenaar, N., & van der Heijden, P. G. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLOS ONE*, 14(3), Article e0213245. <https://doi.org/10.1371/journal.pone.0213245>

Westervelt, E. (2017, August 18). Did a bail reform algorithm contribute to this San Francisco man's murder? *National Public Radio*. Retrieved from <https://www.npr.org/2017/08/18/543976003/did-a-bail-reform-algorithm-contribute-to-this-san-francisco-man-s-murder>

©2020 Eugenie Jackson and Christina Mendoza. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. [Brennan, Dieterich, and Ehret \(2009\)](#) ↵
2. [Lansing \(2012\)](#) ↵
3. [Dieterich, Oliver, and Brennan \(2011\)](#) ↵
4. [Dieterich, Brennan, and Oliver \(2011\)](#) ↵
5. [Farabee et al. \(2010\)](#) ↵
6. [Flores et al. \(2016\)](#) ↵
7. [Reich et al. \(2016\)](#) ↵
8. [Dieterich, Mendoza, and Brennan \(2017\)](#) ↵
9. [Dieterich, Mendoza, Hubbard, Ferro, and Brennan \(2018\)](#) ↵
10. [Rong and Matthews \(2018\)](#) ↵
11. As the authors note, the man accused in that murder had been released after receiving the Public Safety Assessment (PSA) (DeMichele et al., 2018), a short, transparent, summative pretrial risk assessment that is not associated with Northpointe. Ironically, a Los Angeles deputy district attorney is quoted in Westervelt (2017) saying, “It just underscores the problem of having some kind of algorithms that no one really understands making these kinds of determinations,” in reference to the PSA. ↵
12. In fact, Tollenaar and van der Heijden (2019) find the failure to account for nonlinearity is a persistent problem in the literature comparing recidivism models. ↵