Methodological Considerations in Risk Assessment Research

Chapter · July 2016

DOI: 10.1093/acprof:oso/9780199386291.003.0002

Seena Fazel
University of Oxford

Stål Bjørkly
Molde University College

2

Methodological Considerations in Risk Assessment Research

Seena Fazel and Stål Bjørkly

There has been increasing awareness in scientific research of the importance of accounting for possible biases and the need for transparency. This "research on research" has been driven in part by the problems of publication bias in treatment and observational research, and also by the lack of validation for risk factors, associations, and biomarkers in many areas of science, including psychology (Baker, 2015). Furthermore, much research is not applied in clinical practice, sometimes because interventions are not detailed sufficiently in publications to allow for their implementation. This has led some prominent commentators to estimate that more than 90% of all scientific research may be wasted as a consequence (Macleod et al., 2014).

Authorship and Allegiance Bias

In the research literature there are conflicts of interest, many of which are not declared, that may influence what is reported and how. This bias has been found in psychotherapy research in which researchers' self-reported allegiance has been correlated positively with the strength of the preferred psychotherapy's superiority (e.g., Tolin, 2010). It is also worthwhile noting that even statistically significant meta-analyses of clinical trials have been found to have inflated effects as a result of author bias (e.g., Pereira & Ioannidis, 2011). Thus, it is not surprising there is evidence in the field of risk assessment for what is variously called *authorship* or *allegiance bias*—meaning, researchers who were involved in developing an instrument tend to publish more favorable predictive accuracy findings than independent groups (Singh, Grann, & Fazel, 2013). The bias may simply reflect better fidelity with or skills in the use of risk assessment approaches, and appears to be extended to the translators of such tools as well.

Design and Statistical Analysis

Research in risk assessment is not immune to the recognized structural problems in scientific research that have led to various biases. Specifically, a number of problems have been reported, such as poor design, small samples, inconsistent application of risk assessment tools, and incomplete reporting of statistics, some of which have little clinical meaning. A clear example is the overreliance on the receiver–operating curve and the area under the curve (AUC) statistic, which indicates solely whether for any given participant in a trial who has the outcome of interest, the score on a risk assessment instrument is greater than for someone who does not have the outcome of interest. Although risk assessment tools tend to produce similar AUCs, the statistical limitations of this performance indicator mean that two instruments—one that is useful for identifying "high-risk" clients and one that is useful for identifying "low-risk" clients—may produce the same AUC (Singh, 2013). If those two instruments were used in practice for the same purpose, they would lead to very different false-positive or false-negative decisions that would affect public safety and civil rights.

However, even when optimal cutoffs are provided, further statistical information should be included to clarify strengths and limitations, and thus enhance accessibility for clinicians. For instance, actuarial risk assessment tools have been criticized for using risk estimates in the form of proportions for cutoff scores rather than predictive values. Closely related to this, they have also occasionally failed to give clear guidance regarding the interpretation of findings for clinicians and researchers whose local base rates of violence differ significantly from those found in the calibration samples (e.g., Neller & Frederick, 2013). This situation highlights the effect of base rates of violence on validation tests and clinical use of risk assessment tools. It is well known that people tend to overpredict low base rate behaviors (e.g., Kahneman & Tversky, 1985), and this base rate neglect is also present regularly in risk assessments by mental health professionals (e.g., Mills & Kroner, 2006).

There are three implications that arise from this. First, risk assessment instruments should emphasize the importance of considering base rate information in risk judgments. Although some tools are starting to do this (Douglas, Hart, Webster, & Belfrage, 2013), they do not provide more guidance on how to make effective use of this information. Second, research designs should include sufficiently long follow-up so that studies are powered adequately to investigate violent outcomes. In addition, prospective studies with primary end points that are prespecified (ideally by registering protocols) should be conducted. Last, a range of statistics should be used, including those with more clinical utility, such as positive and negative predictive values, true and false positives and negatives, and number needed to detain (Fazel, Singh, Doll, & Grann, 2012). Statistical methods that allow for different base rates and include external validation of models should be used (Fazel et al., 2016).

Contextual and Cultural Issues

Contextual and cultural factors cover a wide range of environmental characteristics that may have an impact on the quality and feasibility of a selected research design. For instance, the population in which the tool was calibrated may not generalize to the clinical group to which one intends to apply an instrument. Much of the research on risk assessment tools has been conducted in white middle-aged men leaving high-security prisons and hospitals in North America (Singh, Grann, & Fazel, 2011), and thus these tools may perform considerably worse in an inner city prison or a medium-security unit in a European country with a different age and ethnic structure. In general, environmental factors are not measured or integrated explicitly in such research. Most risk assessment research assumes similarity of environments, rather than entering environmental variables in the analyses. As a consequence, important bias may run undetected.

Although many studies have been published that examine the validity and reliability of violence risk assessment tools in different countries (Skeem & Monahan, 2011), scant guidance exists on practical considerations when conducting such research. Our intention is not to give a complete overview of all such factors, but rather to highlight some important considerations.

For example, some environmental factors during follow-up may need to be considered for certain populations. Housing, employment, finances, social support, and neighborhood may facilitate or protect against relapse into violence, and they need to be tested empirically if possible. At the same time, there may be little variance in these factors in many discharged patients. The potential relevance of such factors can be illustrated by a hypothetical example of a risk assessment research project. In a predictive validity study of a new risk assessment instrument, 400 patients were assessed at discharge and the raters were blinded with regard to the individual patient's future living situation. Half the patients were discharged without regular supervision to the violent neighborhood from which they originally came (group A), whereas the remaining patients, who originally came from similar living conditions, were transferred to a low-risk neighborhood with daily visits from a psychiatric outreach team and access to drug abuse control, organized employment, and leisure activities (group B). The method for follow-up monitoring of violent incidents was the same for all patients. The result of the test of predictive validity showed high rates of false positives in group B and low rates of false positives in group A. Accordingly, the average assessment accuracy turned out to be just barely above chance. Of course, the example is forced and hypothetical. Nevertheless, it highlights the need to test such factors, if relevant, for possible inclusion in risk assessment tools.

In hospitals, studies have shown the most common victims of violence change over time, with moderating characteristics including whether victims are staff or other patients, as well as patient gender (e.g., Daffern, Mayer, & Martin, 2003). More experience and formal training have been found to decrease staff risk for assault (e.g., Flannery, Staffieri, Hildum, & Walker, 2011). The physical environment of a treatment unit may also influence rates of violence directly through, for example, sensory overload, and indirectly by overcrowding (e.g., Welsh, Bader, & Evans, 2013). There is also a series of relational factors that may differ by psychiatric units and may be relevant. Some factors associated with increasing violence risk are authoritarian and inflexible communication style (e.g., Newbill et al., 2010), lack of consistency in limit-setting situations (e.g., Flannery, Farley, Rego, & Walker, 2007), inadequate response to patients' level of anxiety, provocative staff behavior, high expressed emotional distress in staff, and limited physical and emotional availability of staff (e.g., Cornaggia, Beghi, Pavone, & Barale, 2011; Ross, Quayle, Newman, & Tansey, 2013). There is some evidence, in addition, that the staff-to-patient ratio will affect the quality of follow-up monitoring, with likely underreporting episodes of intrainstitutional violence in units with low staff-to-patient ratios. The ratio may vary from unit to unit within the same hospital, between units providing the same type of services within one country, and between countries.

Research in Routine Clinical Practice

The use of risk assessment tools in routine clinical practice needs further examination. The focus should be on both effectiveness and efficacy. For effectiveness to be demonstrated, there needs to be naturalistic, prospective research on the accuracy of clinicians' assessments in their normal treatment settings. Moncher and Prinz (1991) introduced and defined guidelines for treatment fidelity. They outlined methodological strategies to monitor and enhance the reliability and validity of treatment interventions. In research, there are approaches to increase the likelihood that a study investigates reliably and validly the clinical assessment and/or intervention under scrutiny. However, fidelity measures appear to be almost nonexistent in the research literature on risk assessment of violence. This deficiency may reflect that researchers consider the actual tool to be self-explanatory through user guidelines or manuals. Even if this is true, there is still a need for empirical testing of this assumption, and perhaps even more so for testing the fidelity concerning follow-up of risk management plans.

Closely related to the fidelity topic, another important challenge is to improve follow-up monitoring of violent outcomes. Because a low positive predictive value appears to be a limitation of current risk assessment tools, it is important to ensure all violent acts are reported. Most risk assessment tools do not differentiate types of violence, and outcomes range from verbal threats to homicide. Future studies should provide detailed descriptions of how violence during follow-up was monitored. Reliable measurement of violent threats is important because underreporting appears to be more common for threats than for serious violent acts, although violent crime will remain the most important outcome from a public health perspective.

It can be assumed that intrainstitutional research has the potential for more accurate monitoring than follow-up after discharge to the community. Hence, studies of violence risk judgment in psychiatric facilities are important not only because of their potential of better outcome monitoring, but also because they allow for more accurate measurement of the impact of risk management strategies on predictive accuracy. Another reason is that scrutinizing patient symptoms may enhance risk assessment. For example, Yang and Mulvey (2012) argue for the relevance of assessing subjective experiences of patients for further development of structured assessment methods. They emphasize studies showing that individuals with poor coping strategies for fluctuations in emotional distress have a larger risk of conducting violent acts than persons with similar psychotic symptoms. Accordingly, they recommend examination of the individual's first-person perspective to enhance the predictive validity of dynamic risk factors but could be extended to staff reporting of outcomes. This is relevant methodologically in terms of research design. For example, risk assessments based solely on data obtained from file review may fail to provide information of patients' particular risk factors.

One of the challenges for research in the field is to examine the effects of identifying someone as a high-risk person. The problem is that when a moderate or high risk of violence is communicated, it may lead to interventions to reduce this risk directly or indirectly (through increased surveillance of violence). As mentioned earlier, however, there is little empirical knowledge about this, and currently reviews suggest no difference in those studies using tools retrospectively (after the outcomes have occurred in a case–control design) with those that used them prospectively (cohort designs) (Fazel et al., 2012). Developing a research design based on reliable and repeated measurement of key variables that controls for the possible effects of environmental factors such as risk management strategies would represent a significant step forward. Testing of risk scenarios may possibly contribute to further progress.

Testing of Risk Scenarios

Compared with actuarial tools, structured professional judgment approaches more often include environmental factors, such as living conditions, provision of health and social services, and work and leisure in risk assessments. As part of this approach, scenario-based risk assessment has a central role in the seven-step model of the Historical–Clinical–Risk Management 20 version 3. Underlying the identification of risk scenarios is the question: What might a person do in a given context in the future? Even if a series of risk scenarios could be delineated for each patient, only a few distinct scenarios would be relevant. One scenario to consider is the patient commits violence similar to the most recent violent act. If the patient has committed many violent acts, we could also choose the most frequent or typical one. One possible implication of testing the predictive validity of risk scenarios is that the inclusion of risk management strategies and other contextual factors in this type of risk judgment is a basic premise for its use. To our knowledge, however, there is no research that has tested risk scenarios specifically as predictors of violence. This may be so because risk scenarios are qualitative and not amenable to the methods used currently to test predictive validity. For validity testing of risk scenarios, we need three prediction estimates, all preferably measured on a continuous Likert scale: (a) likelihood of exposure to the actual risk scenario (e.g., 1 point, very low; 5 points, very high), (b) likelihood of violent behavior if exposed (e.g., 1 point, very low; 5 points, very high), and (c) estimated severity of violent act (e.g., 1 point, verbal threats; 5 points, life-threatening violence). Ideally, follow-up monitoring should not only comprise unsuccessful (violence occurred), but also successful (no violence occurred), exposure to risk scenarios.

Two tentative research models for scenario-based risk assessment are suggested: (a) *the ideal model*, which measures prospectively the frequency and severity of exposure to the risk scenario, and (b) *the realistic model*, which measures retrospectively to what extent the risk scenario actually precipitated violence recidivism. So what is the advantage of the scenario-based design compared with validity testing of single items and summary risk judgments? We will not know whether this is a step forward before it has been tested empirically. Methodologically, this design may have some advantages; first, it is a better way to test the possible effect of individual risk factors within the framework of a stress–vulnerability model. This could be done by rating how likely exposure to risk scenarios would be, and subsequently monitoring such exposure during follow-up. In this way, the accuracy of the assessment is tested pertaining to specific risk factors and contexts. Examples of such scenarios are relapse into drug abuse or acute psychosis, contact with criminal peers, and so on. Second, the design allows for testing the predictive validity of a risk estimate concerning the *strength* of the link between the scenario and the violent behavior. An example of this is the use of a case crossover design in which violent behavior is monitored when exposed to the actual scenario versus when there is no such exposure. With this design, each patient acts as his or her own control and the comparison yields effect size estimates of risk contingent on the presence or absence of the risk scenario. Last, the simplicity and clinical relevance of using scenario-based risk assessment as summary judgments may have a strong appeal to clinicians, which may increase the chances of this approach becoming integrated into clinical practice and, hopefully, may also enhance fidelity in empirical research on risk assessment.

Conclusion and Recommendations

We have discussed issues related to evaluating whether a tool can be used in a particular clinical or criminal justice setting. A number of factors should be considered, the first of which is whether the instrument has been tested using standard approaches. This first factor requires using an adequately powered sample with few selection biases, and an examination of independent risk factors using multivariate models. Although some authors have argued that risk factors related only causally to criminality should be included in risk assessment tools, because they will have the strongest risk-reducing effects if addressed (e.g., Coid et al., 2011), such arguments need to be demonstrated empirically, especially because such arguments have not been shown to be valid in other areas of prognostic medicine (e.g. cholesterol and cardiovascular events). Second, the predictive accuracy of the proposed tool needs to be tested using measures of discrimination and calibration. The latter is rarely done. Third, if any new risk factors are included, they need to demonstrate incremental predictive accuracy beyond known risk factors (age, gender, and previous violent crime). Fourth, the tool needs to be tested in a validation sample, and subsequently in independent cohorts, also by researchers without any links to the original developers. Last, an RCT should be undertaken to demonstrate that such a tool improves outcomes. This can be done by comparing the current best practice with an additional tool. Interestingly, despite many hundreds of tools and studies, there is only one such randomized controlled trial (RCT), and this trial found no benefits from administering these tools (Troquete et al., 2013). There are other individual factors that need consideration, including the feasibility and ease of use, its cost, and its time. The costs can be significant if training needs to be undertaken and repeated to use a particular tool. Instruments that can be used widely will benefit from not taking as long to complete as the typical structural clinical judgment tools, which a recent study suggested take 15–16 hours to complete for the first assessment (Viljoen, McLachlan, & Vincent, 2010). Subsequent assessment will be less time-consuming.

In the absence of RCT evidence, clinicians may favor those instruments that can inform management and have dynamic factors that can be tracked over time to determine changes in risk. Some instruments have more dynamic factors than others, but there is no evidence they have better predictive qualities than other tools.

Where does this leave us? Systematic reviews and meta-analyses of observational studies may be relied on in the absence of RCT evidence, but even here there are pitfalls. One of them is the scope of the review and the quality of the statistical approaches used. In a review of reviews, Singh and Fazel (2010) found that many of the reviews in the area of risk assessment included duplicates, did not investigate heterogeneity, and reported clinically uninformative statistics, such as correlation coefficients. Other reviews include authors of a particular instrument, which is considered problematic by methodologists. One influential review compared only head-to-head studies of risk assessment tools (Yang, Wong, & Coid, 2010) and used Cohen's d and ROC AUC (converted to Cohen's d). However, ROC AUC is notoriously insensitive to changes in model performance and may mask important underlying differences in tool's performance. Another review with broader inclusion criteria and used a wider range of metrics reported differences between tools (Singh, Grann, & Fazel, 2011). Although these two reviews differed in their findings with regard to whether risk assessment tools performed comparably, both concluded that such instruments achieve, at best, a moderate level of predictive accuracy.

In practice, decisions on which tool to use may be determined by arbitrary factors, including the success of the marketing of a particular tool. Nevertheless, we suggest some criteria: first, the strength of the evidence for a particular tool, in terms of the quality of the research underpinning it (sample size, transparency of methods, pre-specific protocol, and reporting of key outcomes) and ultimately experimental designs. Second, whether the tool has been validated in a population that is similar to the one for which one wishes to use it. For example, if a population of interest is older violent offenders leaving prison, research validating the tool for this population may be necessary. It is notable that few studies have examined the usefulness of violence risk assessment tools in patient samples with specific mental disorders such as schizophrenia and related psychoses (Singh, Serper, Reinharth, & Fazel, 2011), despite their widespread use in both secure and general psychiatric hospitals. The items in current tools may have poor predictive ability in psychiatric populations (Coid et al., 2011). This is particularly important for criminal history variables as they have strong associations with future offending, such as young age at first violent incident that appears in the HCR-20 and other tools. However, if you take the risk factor, young age at first violent conviction, the largest longitudinal study of risk factors in schizophrenia found that this was not associated with any increased risk of future violent crime—in fact, it had the weakest association of the 15 criminal history factors investigated (Witt, Lichtenstein, & Fazel, 2015). The evidence base is growing every year, and therefore reviews may need to be supplemented by newer primary studies, particularly if they are well conducted and from independent groups. Third, organizations that are focused on quality improvement should institute programs of research and evaluation when a new instrument is introduced, ideally through an RCT, but also through quasi-experimental studies. Such studies should include collecting information on novel risk factors, when hypotheses exist, to consider whether local adaptations to existing tools are needed. For example, in some countries, specific drugs of abuse are associated with offending, or the healthcare system is structured in such a way that the patients with neuropsychiatric problems, such as traumatic brain injury, end up on psychiatric wards. The current research may not have considered these novel factors, and primary research may improve risk assessment in particular countries.

In this chapter we emphasized some different approaches for further progress in violence risk assessment research. We did not present a comprehensive and detailed overview of the many issues involved, but chose to focus on some key issues and those that may be feasible to address. Our main point, however, is to underscore that research on violence risk assessment is in need of innovation (Fazel et al., 2016).

References

Baker, M. (2015). First results from psychology's largest reproducibility test. *Nature.* http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433

Coid, J. W., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D., & Rogers, R. (2011). Most items in structured risk assessment instruments do not predict violence. *Journal of Forensic Psychiatry & Psychology, 22*(1), 3–21.

Cornaggia, C. M., Beghi, M., Pavone, F., & Barale, F. (2011). Aggression in psychiatry wards: A systematic review. *Psychiatry Research, 189,* 10–20.

Daffern, M., Mayer, M., & Martin, T. (2003). A preliminary investigation into patterns of aggression in an Australian forensic psychiatric hospital. *Journal of Forensic Psychiatry and Psychology, 14,* 67–84.

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20^{V3}: Assessing risk for violence: User guide.* Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Fazel, S., Chang, Z., Larsson, H., Långström, N., Lichtenstein, P., Fanshawe, T., Mallett, S. (2016). The prediction of violent reoffending on release from prison: a clinical prediction rule (OxRec). *Lancet Psychiatry* epub 13 April 2016; doi:10.1016/S2215-0366(16)00103-6.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal, 345,* e4692.

Flannery, R., Farley, E., Rego, S., & Walker, A. (2007). Characteristics of staff victims of patient assaults: 15-Year analysis of the Assaulted Staff Action Program (ASAP). *Psychiatric Quarterly, 78,* 25–37.

Flannery, R., Staffieri, A., Hildum, S., & Walker, A. (2011). The violence triad and common single precipitants to psychiatric patient assaults on staff: 16-Year analysis of the Assaulted Staff Program. *Psychiatric Quarterly, 82,* 85–93.

Kahneman, D., & Tversky, A. (1985). Evidential impact of base rates. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). New York, NY: Cambridge University Press.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A, Al-Shahi, S., Chan, A-W, & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *Lancet, 383*(9912), 101–104.

Mills, J. F., & Kroner, D. G. (2006). The effect of base-rate information on the perception of risk for re-offence. *American Journal of Forensic Psychology, 24,* 45–56.

Moncher, F. J., & Prinz, F. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247–266.

Neller, D. J., & Frederick, R. I. (2013). Classification accuracy of actuarial risk assessment instruments. *Behavioral Sciences and the Law, 31,* 141–153.

Newbill, W. A., Marth, D., Coleman, J. C., Menditto, A. A., Carson, S. J., Beck, & N. C. (2010). Direct observational coding of staff who are the victims of assault. *Psychological Services, 7*(3), 177–189.

Pereira, T. V., & Ioannidis, J. P. A. (2011). Statistical significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology, 64,* 1060–1069.

Ross, J., Quayle, E., Newman, E., & Tansey, L. (2013). The impact of psychological therapies on violent behaviour in clinical and forensic settings: A systematic review. *Aggression and Violent Behavior, 18,* 761–773.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8–22.

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: a metareview. *Criminal Justice and Behavior, 37,* 965–988.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499–513.

Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS One, 8*(9), e72484.

Singh, J. P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other psychiatric disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin, 37*(5), 899–912.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20*(1), 38–42.

Tolin, D. F. (2010). Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review, 30,* 710–720.

Troquete, N. A. C., van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: Cluster randomised controlled trial. *British Journal of Psychiatry, 202*(5), 365–371.

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment, 17*(3), 377–395.

Welsh, E., Bader, S., & Evans, S. E. (2013). Situational variables related to aggression in institutional settings. *Aggression and Violent Behavior, 18,* 792–796.

Witt, K, Lichtenstein, P., & Fazel, S. (2015). Improving risk assessment in schizophrenia: epidemiological investigation of criminal history factors. *British Journal of Psychiatry, 206*(5), 424–430.

Yang, S., & Mulvey, E. P. (2012). Violence risk: Re-defining variables from the first-person perspective. *Aggression and Violent Behavior, 17,* 198–207.

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*(5), 740–767.