


# Empirical attrition modelling and discrimination: Balancing validity and group differences

Andrew B. Speer 

Department of Psychology, Wayne State University, Detroit, Michigan, USA

## Correspondence

Andrew B. Speer, Department of Psychology, Wayne State University, 5057 Woodward Ave (room 8402.24), Detroit, MI 48202, USA.  
Email: [speerworking@gmail.com](mailto:speerworking@gmail.com)

## Abstract

Attrition models combine variables into statistical algorithms to understand and predict employee turnover. People analytics teams and external vendors use attrition models to offer insights and to develop organisational interventions. However, if attrition models or other data-driven models inform employment decisions, model scores may then be subjected to civil rights laws and diversity concerns resulting from group differences in scores. This paper discusses adverse impact when building attrition models, outlining how researchers test for adverse impact in this context, strategies to reduce group differences and how attrition modelling and other human resources 'big data' predictions fit within larger validity frameworks. Procedures were applied to field data in an applied demonstration of an attrition model with disparate impact. Model revisions resulted in adverse impact reductions while simultaneously maintaining model validity. Collectively, this paper provides timely attention to important aspects of the people analytics, turnover and legal domains.

## KEYWORDS

adverse impact, attrition modelling, big data, discrimination, HR analytics, machine learning, people analytics, turnover

**Practitioner notes****What is currently known?**

- Attrition models combine variables into statistical algorithms to understand and predict employee turnover.
- However, if attrition scores or other data-driven models inform employment decisions, such models may then be subjected to civil rights laws if group differences exist.

**What this paper adds?**

- This paper describes in detail how practitioners should test attrition models for adverse impact, how to reduce adverse impact and how attrition models fit within larger validity frameworks.
- This paper also provides an empirical demonstration of building an attrition model using field data. The developed model was valid but had adverse impact. A search for alternative models was performed by revising the algorithm to reduce adverse impact while maintaining validity.

**Implications for practitioners**

- The procedures described in this study are directly applicable to practitioners who seek to build attrition models or other data-driven models by helping them adhere to legal guidelines and best practices.
- Attrition models should be tested for adverse impact, and revisions to attrition models should be made in efforts to reduce adverse impact while maintaining validity.

## 1 | INTRODUCTION

People analytics involve the analysis of organisational data to understand employee phenomena and improve business outcomes (Marler & Boudreau, 2017). The current paper focuses on the use of available organisation data to statistically (e.g. via modern prediction methods/machine learning, ML) predict employee turnover. Such work has recently been referred to as attrition modelling (Speer et al., 2019), which is formally defined as applying statistical algorithms to understand and predict employee turnover (Speer et al., 2019). The data inputs for such algorithms include constructs assessed using traditional surveys, as well as more novel and dynamic data, often coming from disparate data systems (e.g. time and tracking information, communication meta-data). Attrition modelling produces turnover probabilities and expected turnover rates for groups or individuals, which in turn feeds into organisational decision making, workforce planning and various strategic decisions. Unlike generalised inquiries into turnover, attrition modelling is primarily data-driven, and it is a practice well-suited for people analytics teams. Internal human resources (HR) analytics teams have access to data to predict turnover and the outputs can be easily applied as positive contributions to the organisation.

Despite the potential benefits of attrition modelling, there are few existing empirical studies that outline the decisions involved and analyses used in attrition modelling. Furthermore, there may be obstacles and risks to attrition modelling implementation. One particular risk, and the one that is the focus of this paper, is that attrition models have the potential to differentially impact protected demographic groups; when this occurs, attrition models or other HR 'big data' interventions would then be subject to major civil rights laws. More specifically, and as mentioned by Obenauer (2019) and Castille and Castille (2019), any employment decisions based upon attrition models or other data driven HR predictions can be subject to disparate treatment and adverse impact scrutiny. For example, if predicted turnover scores were significantly different between protected groups (e.g. genders), this

could result in adverse impact when those scores are used to determine resulting employment decisions (e.g. entry into coaching programs, placement on improvement plans, pay changes). The presence of group differences in attrition scores thus poses legal risk if used to influence employee decisions, resulting in questions regarding how to test for discrimination in this context, and if found, what procedures can be implemented to continue utilizing attrition models. It also raises the question of whether certain attrition models are counterproductive to an organisation's diversity goals.

The current paper expands upon recent commentaries on attrition modelling (e.g. Castille & Castille, 2019; Speer et al., 2019) by providing deeper discussion and an applied demonstration of the complexities involved when dealing with group differences in attrition scores, including the need to test for alternative models with less adverse impact but similar validity. This paper is applicable to both researchers and practitioners, as it examines the conditions and strategies for discrimination testing and for establishing validity evidence for attrition models. The principles can also be extended to other big data HR applications too (e.g. high potential identification). Although Castille and Castille (2019) broadly discuss group differences with attrition models, the current paper provides a much more detailed look at these issues at a level of specificity that is necessary for practitioners to responsibly perform attrition modelling. Such a focus is important both in terms of maintaining legal protection when using attrition models and other data-driven HR models, but also to ensuring that companies are achieving their diversity goals. Additionally, this study also contributes to the literature by illustrating the use of attrition modelling using actual field data. Attrition modelling (and ML at large) is increasingly being applied in the organisational sciences, and this study is an applied demonstration of how attrition modelling can be conducted. Relatedly, this paper serves as an illustration of how big data methods must be implemented with care and adhere to common concerns regarding fairness and validity. There is a fervour around big data applications in HR, but this paper illustrates that implementation of such methods is not a simple matter; instead, implementation can have profound and unexpected impacts. Thus, this paper connects newer HR methods to more traditional topics in the organisational sciences.<sup>1</sup>

Taken together, this paper is intended for practitioners who work in attrition modelling and people analytics, and to academics who research turnover, big data and newer HR methods, and employment discrimination. This paper begins by providing an overview of attrition modelling, followed by a discussion of how attrition modelling can result in unequal attrition scores across demographic groups. This discussion details disparate impact and treatment within attrition modelling, therefore connecting this practice to major civil rights law in countries where people analytics is practiced and civil rights protections exist (e.g. United States, European Union). Finally, this paper connects adverse impact and attrition models to a validity framework used when evaluating employment procedures. Then, the paper outlines a step-by-step procedure to address group differences within attrition modelling or other data-driven HR predictions. This procedure is applied to actual field data to illustrate important decisions researchers make when performing attrition modelling.

## 1.1 | Brief overview of attrition modelling

Attrition models combine variables that predict turnover into statistical algorithms that then estimate the probability of employee turnover within a given timeframe, or at a specific timepoint.<sup>2</sup> The variables typically derive from internal company databases such as human resource management systems (HRISs) but may also combine other interesting data (e.g. communication metadata). The number of variables often can be quite large. Conway and Frick (2017) utilised 87 variables to create turnover probabilities for a large population of workers. Likewise, Speer et al. (2019) describe an example where a company built an internal attrition model from dozens of company variables (e.g. job performance, job changes, manager changes). These models often utilise ML to improve prediction in new samples (e.g. Es-Sabahi & Deluca, 2017; Rosett & Leinweber, 2017).

The formed attrition estimates can then serve a number of purposes, including use for pre-employment selection (Gibson et al., 2019; Strickland, 2005), to validate and develop training initiatives (McCloy et al., 2016; Strickland, 2005), to facilitate workforce planning discussions with specific parts of the company (Speer et al., 2019), to create ad hoc programs to reduce attrition (Strickland, 2005) and a variety of other HR purposes generally aimed at understanding and impacting employee turnover. The work is conducted both internally and by external vendors as well. For example, HR software companies currently offer features that include projected group-level turnover estimates within HR dashboards, as well as risk projections for individual employees. These are often accompanied by in-depth studies into the root causes of turnover, which then facilitate turnover interventions. Thus, attrition models serve various strategic HR purposes.

## 1.2 | Discrimination in attrition modelling

As noted by Castille and Castille (2019), attrition algorithms, if used to determine employment decisions, may be subject to civil rights laws. Within the United States (US), employment decisions, as originally outlined in the Civil Rights Act of 1964 and since codified via case law and other legislation, cover a swath of employment practices and conditions of employment such as hiring, promoting, compensating, terminating and other general conditions or privileges of employment. Protections occur according to race, colour, religion, sex and national origin, as well as other classes via other major US legislations (e.g. Americans with Disabilities Act, Civil Rights Act 1991, Age Discrimination in Employment Act, Equal Pay Act, Pregnancy Discrimination Act); many of these demographic variables are similarly protected in other countries (e.g. those in the European Union).<sup>3</sup> There is a long history of litigation involving employment decisions that could be influenced by attrition modelling or other data-driven HR predictions, including pay (Doverspike et al., 2019), promotions (Siskin & Schmidt, 2019), terminations or reductions in force (Deere & Pearce, 2019), and other HR processes such as succession planning (Lundquist et al., 2019). If attrition modelling or other data-driven HR predictions affect any of these decisions or other employment outcomes, then attrition models would therefore be scrutinised for potential discrimination against protected groups. This is likely to occur when attrition models directly inform employment outcomes (e.g. when used as one determiner of who receives training or pay), but also when attrition modelling has more distal and indirect effects on long term systemic discrimination (for a non-attrition example, see *International Brotherhood of Teamsters vs. United States*, 1977).

Discrimination is traditionally differentiated as either disparate treatment or adverse impact (i.e. disparate impact). Disparate treatment involves explicitly using protected group status as a component of decision making. Creators of attrition models may not intend for model scores to be used for employment decisions, instead using such scores as a research tool to better understand organisational turnover trends. This may lead some researchers to incorporate group-level variables such as race, gender and age directly into attrition models. However, if such factors are incorporated into those models and those models are ever used to make employment decisions, this is illegal. Thus, protected characteristic variables should not be used in attrition modelling or other data-driven HR models used for decisions about individual employees.

Companies may also unintentionally run afoul even if protected characteristics are not included in attrition models or other data-driven HR predictions. When attrition scores are used to facilitate employment decisions, this would occur if the decisions influenced by attrition scores differ by protected group. Such an effect would constitute adverse impact, or when facially-neutral procedures produce discriminatory effects on protected groups. Many variables that are commonly related to turnover propensity covary with protected class membership (e.g. job performance, zip code). Obenauer (2019) also notes that demographic groups have different attrition rates in the US, thus directly implicating the targeted outcome variable. Furthermore, when attrition algorithms are based on big data and lots of variables are combined via ML, it is likely that captured variance will be related to some

protected group characteristic (Castille & Castille, 2019). Thus, attrition models run the risk of exhibiting group differences in expected termination rates even if protected class membership is not included within models.

One can imagine just how easily these issues could manifest into legal issues with attrition scores or other data-driven HR predictions. For example, a well-known HR software company offers dashboards that include turnover propensity estimates at the group and individual levels, along with predictor-level drivers of turnover (e.g. commute distance). In at least one instance, these dashboards were implemented without any training for the end-users. In these cases, HR managers might fail to understand the accuracy of the models and implement initiatives without solid local evidence of validity, or they may incorrectly utilise the output (Angrave et al., 2016). Results might be shared with constituents who should not be privy to such data, resulting in possible behavioural implications for turnover-risk employees. Or, turnover correlates might be haphazardly leveraged for employee selection and promotion.

There are no known disparate treatment or impact court cases that involve attrition modelling, and there is very little established legal discourse regarding other 'big data' HR predictions. However, because such scores can be used to facilitate employment decisions related to pay, promotions, terminations or succession planning, all of which have discrimination case law and legal precedence (Deere & Pearce, 2019; Doverspike et al., 2019; Lundquist et al., 2019; Siskin & Schmidt, 2019), this means attrition scores and other big data HR models can be subject to legal scrutiny. Just because case law does not yet exist does not mean that settlements have not been made, or that there is no future risk for litigation. Regarding big data HR methods more broadly, experts believe related court cases are inevitable (Lewis, 2019). Rather than be reactive, practitioners would benefit from steps they can take to be proactive in preventing legal risk.

### 1.3 | Attrition model discrimination analysis

In this section, I outline steps to account for group differences in attrition scores as recommended by Castille and Castille (2019). However, the framework can easily be applied to any data-driven HR model. (1) First, the attrition model is built and policies are established regarding how model scores are to be used. If used for decision making, this model should be built without any protected class variables to avoid disparate treatment. Although not mentioned by Castille and Castille, it would be appropriate to examine the validity of the developed model at this stage also, which will be elaborated upon later.

(2) After the model and policies are established, the model is examined for adverse impact. What complicates matters for attrition modelling is that (a) attrition models may not be used in a binary fashion and (b) the employment decision may shift such that for some decisions employers may target those with low attrition risk, whereas for other decisions employers may target those with high attrition risk. For example, those with low attrition risk might be entered into desirable and expensive training programs under the assumption that training would be best utilised for employees likely to stay at the company (e.g. if the cost of entry into the organisation is high). On the other hand, a company might instead operate in the opposite fashion; employers might try to dissuade employees from leaving (and especially top-performers) by entering those more likely to terminate into training programs in hopes their likelihood of retention increases, or by offering pay increases. Thus, the directionality of what is coded as the 'decision' can change based on how employment decisions are made. This then affects which group is 'favoured' by the attrition scores. The attrition scores might also be used in ad hoc fashion (e.g. managers addressing employee concerns on their own), which might make it difficult for plaintiffs to identify attrition scores as the practice causing discrimination. These unique features make it challenging to establish majority groups and generally determine how to approach adverse impact analyses in this context.

There is a rich literature on traditional adverse impact statistics (Morris & Dunleavy, 2019), and it is generally recommended that adverse impact is tested for both practical and statistical significance (Murphy & Jacobs, 2012). Practical significance historically includes 4/5ths statistics (Equal Employment Opportunity Commission

et al., 1978) and would be performed if the continuous attrition scores are dichotomised. Of course, issues with the 4/5ths rule are widely known, and these include susceptibility to type 1 and type II errors, the effects of decision rate magnitude on the likelihood of violation and the general arbitrary value of 80% as a cutoff (Oswald et al., 2019). In the case of attrition scores, and especially when no strict attrition dichotomy is utilised, researchers might instead focus on standardised mean differences or percent variance explained for practical significance estimates of continuous attrition predictions (Murphy & Jacobs, 2012). Benchmarks for what constitutes as small, medium and large effects are generally well accepted within the social and organisational sciences (J. Cohen, 1988; Murphy & Jacobs, 2012). Likewise, these can easily be paired with inferential tests (t-tests, F-tests) to determine whether there are significantly different attrition predictions for protected groups. Such an approach is recommended irrespective of whether the predicted scores are dichotomised or not.

When models facilitate clearly differentiated decisions based on the model predictions, statistical significance should also be calculated based on comparisons of decision rates. Statistical significance testing has become increasingly relied upon by courts (Morris & Dunleavy, 2019) and such tests calculate whether decision rates likely differ between groups by a value greater than zero. There are a variety of statistical tests for these purposes, and the reader is referred to Morris and Dunleavy for guidance on calculations. This study will apply the Z-test for the difference in selection rates, which is commonly and incorrectly<sup>4</sup> referred to as the '2 standard deviation rule'.

In the (3) third step suggested by Castille and Castille (2019), researchers act upon the findings from step 2. If adverse impact is found, researchers have several options, though these are not thoroughly explicated in Castille and Castille's paper. In line with adverse impact tradition, the company could establish whether model interpretations are valid, thus insulating from legal risk in the presence of adverse impact. This process will be described in the next section. Then, if the procedure is found to be valid, or if the company simply wishes to reduce group differences to promote diversity or further minimise legal risk,<sup>5</sup> the company would consider alternative decision procedures. This might involve creating an entirely new but equally valid model for decision making (i.e. alternative procedure). Or, it might involve tweaking the existing model to reduce differences in scores between protected groups.

King and Mrkonich (2016) describe how algorithm adjustment (removal of variables, changing variable weights) might be performed as an alternative when algorithms have discriminatory impact. For example, researchers might identify variables with moderate-to-large standardised mean differences between groups and remove those variables. However, application of this in practice might create a 'whack-a-mole' conundrum where adjustments to variable weighting to reduce discrimination for one protected group unintentionally results in new discrimination against another protected group (King & Mrkonich, 2016). Furthermore, depending on the complexity of the attrition algorithm, it may be challenging to discern whether a variable helps or hinders a certain group in terms of model outputs. Variable suppression and directional ambiguity of variable weighting within complex algorithms (e.g. neural networks) make it challenging to determine which variables to remove. In these cases, researchers may need to rely on theoretical considerations and zero-order relationships between each individual predictor and the predicted outcome. Another option is to use Pareto optimisation algorithms that derive weights to maximise prediction while minimizing adverse impact (De Corte et al., 2007), or ML analogues for neural networks (e.g. adversarial debiasing, Wadsworth et al., 2018).

### 1.3.1 | Incorporating validity evidence

Employment decisions are legally supported when those decisions are job or business-related. Drawing from the large case history for employee hiring, simply having group differences in decision rates does not preclude use of an employment procedure if sufficient validity evidence exists (Principles, 2018). This extends to other employment decisions as well, such as performance reviews (Martin et al., 2000; Werner & Bolino, 1997), promotions (Siskin &

Schmidt, 2019) and terminations (Deere & Pearce, 2019). Similarly, if companies can establish that attrition models predict turnover, policies based on those models should theoretically be supported even when adverse impact is present. Reduction of turnover is by itself a legitimate business pursuit, as turnover has traditionally been viewed as a valid concern across employers and irrespective of job analysis (Principles, 2018). However, attrition scores can be used in many ways. Thus, there should be consistency and business logic in how particular interventions or employment decisions are linked to attrition probabilities. Assuming that exists, the matter then becomes whether attrition scores are valid as predictors of turnover.

Contemporary practice adheres to the unitarian view of validity, in line with the Standards for Educational & Psychological Testing (2014), such that all strategies and sources of validity evidence (e.g. content, associations with other variables) serve to establish overall construct validity. Nonetheless, it is worthwhile to consider how the different sources of validity evidence might be applied in the attrition modelling context. Given the multidimensional and highly empirical nature of attrition models, as well as the theoretical complexities of turnover decisions (e.g. Hom et al., 2012), content-oriented and construct-oriented validation strategies are likely to pose challenges for attrition modelling. As such, criterion-related validity will be the focus of this paper, represented by the degree to which attrition estimates relate to actual turnover. Attrition scores should be both significantly (statistically) and practically related to turnover. Speer et al. (2019) suggested common practical correlational benchmarks (J. Cohen, 1988 of 0.1, 0.3, and 0.5 as small, moderate, and large) to infer the practical magnitude of turnover relationship. When the algorithm is used to make dichotomous employment decisions, they also suggested evaluating model performance based on area under the curve (AUC) statistics, with benchmarks taken from Rice and Harris (2005) where 0.56, 0.67 and 0.79 represent small, moderate and large effects (these are the correlational equivalents to Cohen's benchmarks). While these values help establish rules of thumb researchers can abide by, the necessary relationship strength within the turnover context specifically is debatable.

It should be noted that attrition scores will generally exhibit strong correlations with turnover, given many variables are often incorporated within attrition models (Speer et al., 2019). What is likely then more relevant from a legal perspective is whether a reasonable alternative model with less adverse impact is similarly valid in predicting turnover. Possible alternative procedures that are equally valid but result in less adverse impact must be considered in the presence of adverse impact. Although adverse impact theory typically pertains to employee hiring and promotions, as well as pay decisions (Doverspike et al., 2019), the same alternative practices logic can be applied to attrition modelling or other data-driven HR models. This means that validity should be established for attrition models and that researchers should engage in the steps described throughout this paper to evaluate alternative attrition models if adverse impact exists. If a model is found to have adverse impact but a similarly valid model does not have adverse impact, researchers should use that alternative model. Thus, it is important that researchers try to balance both validity and adverse impact when building attrition models and other data-driven HR predictions, and this practice will be described using an empirical demonstration of attrition modelling within this paper.

## 1.4 | Recap and introduction to empirical demonstration

This paper has thus far outlined important considerations when dealing with group differences in attrition scores and other data-driven HR models. However, this discourse is best solidified with actual empirical demonstration. The current paper provides an empirical example of attrition modelling, including application of ML to form attrition estimates. After developing an attrition model for a sample of 894 call centre employees, validity and adverse impact analyses were performed. Then the attrition model was revised in attempts to reduce adverse impact while maintaining model validity.

## 2 | METHODS

### 2.1 | Participants

Data came from 894 call centre employees who worked for a telecommunications company. Employees were responsible for answering customer inquiry calls. During these calls, employees provided service support while simultaneously seeking to obtain additional sales. Turnover was conceptualised as a dichotomous variable as to whether these employees left the company within a 6-month period. At the start of the time period, employees varied in tenure ranging from 11 days to just less than 3 years. Forty-five percent of the samples were White, 17% Hispanic and 32% Black. All other races were grouped as 'other'. Fifty-nine percent were male and 41% female. Average age was 32.6, with 21% being greater or equal to 40 years of age.

### 2.2 | Measures

HRIS variables and performance variables were included within the attrition model. The *HRIS variables* included variables available via the company's HRIS. These existed for all employees and were used within the attrition modelling in some capacity. In addition to demographics (race, gender, age), variables were identified that had theoretical links to turnover. The included variables were job tenure, whether the employee was hired internally, whether the employee changed departments within the company to take the current position, department as a categorical predictor, education level (ordinally scored), whether the employee experienced a change in manager the past year, pay, recent pay change and work location as a categorical predictor. HRIS variables are readily available in organisations, making them convenient variables within attrition models, even if they may not be the strongest predictors. Unfortunately, survey-based measures of attitudes, which are commonly very predictor of turnover (Rubenstein et al., 2018), were not available.

*Performance variables* also existed for the call centre employees, who were regularly tracked according to six metrics, and these were also included within the attrition models: (a) total units sold, (b) sales efficiency, (c) sales commission, (d) first call resolution, (e) customer satisfaction ratings and (f) number of calls handled. For all six metrics, performance was reported monthly and the dataset contained an average monthly performance score for employees. All variables were included within the attrition modelling. Furthermore, employees were annually rated by their supervisors regarding their job performance. Ratings were made on a 1–5 scale where 5 = exceptional performance. Because employees who were just hired but terminated early into the 6-month period did not have performance data for some performance measures, missing performance data were imputed using the multivariate imputation via chained equations package (i.e. 'mice') from R (R Core Development Team, 2007). Missing data were imputed using all variables in the dataset to create a single file with full values. The inter-correlations among continuous variables can be found in Table 1. How the HRIS and performance variables were used in the attrition modelling process can be found in Section 2.3.

#### 2.2.1 | Turnover

Turnover was coded into two general categories within the HRIS: voluntary and involuntary. Voluntary turnover accounted for employee-initiated resignations. Involuntary turnover occurred when the company discharged employees for poor performance. Of the 894 employees, 32% ( $N = 285$ ) turned over within 6 months, with 20% ( $N = 179$ ) being voluntary terminations and 12% ( $N = 106$ ) being involuntary terminations.



TABLE 1 Correlations among turnover and continuous predictor variables

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Turnover																	
2. Job tenure	−0.11																
3. Gender	−0.07	0.09															
4. Age	−0.04	0.06	−0.02														
5. Department change	−0.16	0.56	0.06	0.03													
6. Internal candidate	−0.14	0.30	0.07	−0.04	0.45												
7. Education level	0.03	0.01	0.03	0.09	0.00	−0.07											
8. Supervisor change	−0.11	0.59	0.06	0.05	0.46	0.22	−0.01										
9. Pay	−0.17	0.37	0.07	0.08	0.39	−0.01	0.10	0.27									
10. Pay change	−0.15	0.27	0.06	0.07	0.44	0.20	0.02	0.28	0.70								
11. Sales commission	−0.16	0.30	0.14	−0.06	0.08	−0.07	0.04	0.16	0.13	0.06							
12. First call resolution	−0.02	0.25	0.04	−0.06	0.13	0.07	0.04	0.23	0.11	0.10	0.05						
13. Calls handled	−0.09	0.48	0.06	0.02	0.29	0.25	−0.06	0.49	0.01	0.19	0.21	0.16					
14. Customer satisfaction	−0.05	0.00	−0.01	0.05	−0.02	0.00	0.03	−0.02	0.07	0.08	0.00	0.22	0.01				
15. Units sold	−0.18	0.00	0.13	0.04	−0.01	0.01	−0.07	0.05	−0.15	−0.02	0.15	−0.03	0.31	0.05			
16. Sales efficiency	−0.24	−0.02	0.07	−0.02	0.03	0.09	−0.12	0.03	−0.15	−0.02	0.05	−0.04	0.15	0.10	0.73		
17. Performance rating	−0.09	0.35	0.04	0.08	0.28	0.09	0.06	0.28	0.28	0.28	0.33	0.13	0.20	0.06	−0.01	−0.02	

Note:  $N = 894$ . Categorical variables of race, location and department are not included. For gender, men are coded 1. Correlations  $\geq 0.07$  are  $p < 0.05$ , and values  $\geq 0.09$  are  $p < 0.01$ .

### 2.3 | Formation of attrition model

Many analytical methods can be applied for attrition modelling. A limited review of these can be found in Speer et al. (2019), and coverage of a larger array of ML methods can be found in Kuhn and Johnson (2013). Putka et al. (2018) also provide a user friendly review of these methods for organisational scientists. For this study, the analysis purpose was to identify which employees terminated within a timeframe of 6 months, allowing for a static analytical approach. Ultimately, random forests (Breiman, 2001; Kuhn & Johnson, 2013; Putka et al., 2018) was chosen after considering other possible ML algorithms (e.g. logistic regression, elastic net regression, deep neural networks). Random forests was chosen because it generally results in strong cross-validation, can capture interactions and quadratic effects (which may be relevant to turnover prediction, e.g. Becker & Cropanzano, 2011), does not require prohibitively large sample sizes, and is relatively simple to implement compared to more complicated models, such as deep neural networks. Per request of the reviewers, models were also fit for other ML model types, including logistic regression, elastic net logistic regression (Zhou & Hastie, 2005) and tuned deep neural networks (Goodfellow et al., 2016).<sup>6</sup> Random forests exhibited the best performance when predicting overall turnover, with validity coefficients being 8% higher than logistic regression, 8% higher than elastic net logistic regression and 1% higher than a deep neural network. Given that algorithm choice is dependent upon the nature of the outcome variable(s), the nature of the predictors and the sample size, these results do not mean that random forests is a superior algorithm for predicting turnover. Instead, it simply means that for this particular dataset, random forests was marginally better than other algorithms.

Random forests create a large number (usually several hundred to several thousand) of individual trees to predict an outcome of interest (either categorical or continuous). The trees split predictor scores based on whether the variable split differentiates respondents in terms of the outcome variable. This naturally models non-linear predictor effects and interactions that exist in the data. Whereas a single tree will utilise all variables and use all respondents within the analysis, random forests intentionally reduces each tree's individual predictive power in the calibration sample by randomly sampling from the pool of predictor variables. It also randomly samples from the calibration sample's respondents to create each tree. These aspects reduce the performance of each individual tree, but when those trees are aggregated into an ensemble (i.e. composite of all trees), the result is usually a highly predictive model that minimises variance in new settings.

To create predicted attrition scores using random forests, a nested five-split cross-validation procedure was applied (see Kuhn & Johnson, 2013). This type of procedure is common in ML, and it works by splitting the total sample ( $N = 894$ ) into random 80% calibration samples ( $N = 715$  to  $716$  employees) and 20% holdout samples ( $N = 178$  to  $179$  employees) five times; each holdout sample was independent such that a respondent was only included in a holdout sample once. Attrition algorithms are built using the calibration samples and cross-validated on the holdout samples, providing a test of how well the developed turnover predictions relate to turnover in independent settings. All analyses were done in R (R Core Development Team, 2007), with random forests models estimated using the 'randomForest' package. Optimal hyper-parameters for the number of trees and the number of sampled variables per node were determined using nested k-folds cross-validation within the first 80% calibration sample, resulting in 2500 trees and six variables per node. Cross-validation was done a total of five times, or once for each of the holdout samples, making for a total holdout sample  $N$  of 894 (i.e. total sample  $N$ ). The created attrition estimates were saved for each holdout sample so that each respondent had attrition scores. These were then tested for adverse impact.

Multivariate categorical prediction was performed such that random forests was used to estimate the probability of the three potential turnover outcomes in this study: stay (no-turnover), voluntary turnover and involuntary turnover. Variable selection is naturally conducted within random forests by only splitting variables when that split meaningfully reduces model error. Thus, all HRIS and performance variables were included in the modeling,<sup>7</sup> and at each tree node six variables were randomly sampled (per the hyper-parameter estimated previously); the variable split that reduced error the most at that part of the tree was then made. After models were

run, estimates for voluntary and involuntary turnover were added together to create an estimate of overall turnover probability.

For this study, three sets of attrition models were created using the aforementioned approach. First, a model labelled (a) 'Full Attrition Scores' was formed. This included all predictor variables, including demographic variables. Per the commentary in this paper, this would be an ill-advised model. Nonetheless, it is worth examining how inclusion of demographic variables affects model performance. Second, 'Operational Attrition Scores' were created by including all variables except protected group characteristics. This would be the initial model formed under step 1 of Castille and Castille's (2019) procedure. Finally, (c) 'Revised Attrition Scores' were created. Results from the second set of attrition scores were examined and variables with large group differences were removed in efforts to reduce adverse impact, in line with step 3.

Lastly, to simulate usage of attrition models for decision making, a 50% cutoff was set for the attrition scores. This value of 50% is arbitrarily set, and in practice researchers might adopt other values, though there is no consistently firm strategy for doing so (Speer et al., 2019). A 50% cutoff has inherent meaning, such that crossing this threshold means an employee has a greater likelihood of turning over rather than staying. This study operated under the assumption that those with higher retention scores (i.e. lower attrition estimates) would be favourably impacted in terms of employment, therefore coding employees with a less than 50% turnover probability as (1) and those with a higher than 50% turnover probability as (0). This cutoff was utilised for adverse impact analyses. Note that continuous attrition probability scores were also examined.

As mentioned in Section 1, whether desirable employment decisions are made for those with high versus low attrition scores may differ by context. Those with low attrition risk might be viewed as having better return-on-investment if the company plans to provide limited training or other opportunities (e.g. bonuses), which would reflect the coding used here. Because performance correlated negatively with turnover (Table 1), low risk employees would also be more desirable for the employer in this demonstration. On the other hand, employees with high attrition risk might be targeted to dissuade them from leaving the company (e.g. giving them pay increases). This might occur if the company wishes to retain all employees. Both strategies make sense in certain contexts. Because courts now generally identify the majority comparison group as the group with highest favourable decision rate (D. Cohen et al., 2019), this would have notable effects on which groups are adversely affected. As will be seen, Whites, young employees and men had lower attrition scores in this study and were therefore treated as the majority groups (i.e. referent groups). If instead those with high attrition scores received favourable employee decisions, Whites, young employees and men would then be those adversely affected. This would subsequently work against diversity goals. These features make adverse impact analyses much more complicated when dealing with attrition scores or other HR data predictions, and thus users need to clearly understand which groups to code as advantaged versus disadvantaged based on how the attrition model is used.

## 3 | RESULTS

### 3.1 | Step 1: Model creation and examination of validity

Correlations between predictor variables and turnover can be found in Table 1. Table 2 displays validity coefficients and AUC values for the created attrition scores. Attention will be devoted to prediction of total turnover within this paper, but readers can refer to the tables to examine relationships for other forms of turnover.

As seen in Table 2, Operational Attrition Scores (excluding demographic variables) exhibited a strong and significant relationship with turnover ( $r = 0.48$ ,  $p < 0.01$ ,  $AUC = 0.79$ ). Thus, there was validity evidence in support of scores. As a comparison, although inclusion of demographic variables resulted in a slight uptick in prediction for the Full Attrition Scores ( $r = 0.49$ ,  $AUC = 0.79$ ), this difference is not practically nor statistically significant (Steiger  $Z = 1.40$ ,  $p = ns$ ).

TABLE 2 Relationships between attrition scores and holdout turnover

	Turnover		Voluntary turnover		Involuntary turnover	
	<i>r</i>	AUC	<i>r</i>	AUC	<i>r</i>	AUC
Full attrition model						
Overall	0.49	0.79	0.28	0.70	0.36	0.80
Voluntary	0.36	0.73	0.27	0.69	0.18	0.68
Involuntary	0.46	0.77	0.18	0.66	0.44	0.82
Operational attrition model						
Overall	0.48	0.79	0.27	0.69	0.36	0.80
Voluntary	0.35	0.72	0.26	0.68	0.19	0.68
Involuntary	0.46	0.77	0.18	0.65	0.45	0.83
Revised attrition model						
Overall	0.46	0.78	0.25	0.69	0.36	0.80
Voluntary	0.33	0.71	0.22	0.67	0.20	0.68
Involuntary	0.44	0.76	0.18	0.64	0.41	0.82

Note:  $N = 894$ . All values are  $p < 0.01$ .  $r$  = correlation and AUC = area under the curve. Full attrition model included all variables plus protected characteristics (race, gender, age). The operational attrition model included all variables excluding protected demographic variables. The revised attrition model removed variables with larger group differences (sales commission, work location, units sold, tenure). For all models, probabilities were calculated that individual employees would voluntarily or involuntarily terminate. These were summed together to create overall turnover estimates.

### 3.2 | Step 2: Initial adverse impact analysis

Table 3 contains results from the adverse impact analyses based on overall turnover probabilities.<sup>8</sup> Whites, men and younger employees had lower predicted attrition, making them majority groups per this study's operationalisation. Operational Attrition Scores exhibited weak to moderate standardised mean differences that were significant when comparing Whites to Blacks ( $d = -0.28$ ,  $p < 0.01$ ), Whites to Hispanics ( $d = -0.21$ ,  $p < 0.05$ ), men to women ( $d = -0.25$ ,  $p < 0.01$ ) and young to old ( $d = -0.19$ ,  $p < 0.05$ ). When considering adverse impact in terms of decision rates, neither the 4/5ths rule nor the test for statistical differences in selection rates were violated when comparing Whites to Blacks (adverse impact ratio, AIR = 0.92,  $Z = 1.82$ ,  $p = \text{ns}$ ), Whites to Hispanics (AIR = 0.95,  $Z = 1.06$ ,  $p = \text{ns}$ ) and young to old (AIR = 0.97,  $Z = 0.73$ ,  $p = \text{ns}$ ). However, there were significantly different selection rates against women (AIR = 0.90,  $Z = 2.70$ ,  $p < 0.01$ ). While this was not a practically large difference, it still may be cause for concern, given that some courts rely on statistical significance tests as evidence for adverse impact (e.g. Eissenstat, 2016). In comparison, when demographic variables were explicitly included within the attrition modelling (Full Attrition Scores) adverse impact generally increased, with practical and significant effects occurring for all comparisons besides age.

### 3.3 | Step 3: Creating revised Operational Attrition Scores

Given these findings, Step 3 of the analysis involved algorithm revision to reduce adverse impact. For the sake of simplicity, variables were removed in a univariate fashion in line with Castille and Castille's (2019) recommendations. To do this, each predictor was examined for possible removal from the Operational Attrition Model.

TABLE 3 Adverse impact statistics for attrition scores

	Continuous attrition scores <i>d</i>	Turnover decision	
		AIR	Z
Operational attrition model			
White-Black	−0.28**	0.92	1.82
White-Hispanic	−0.21*	0.95	1.06
Men-women	−0.25**	0.90	2.70**
Young-old	−0.19*	0.97	0.73
Full attrition model			
White-Black	−0.36**	0.89	2.80**
White-Hispanic	−0.25**	0.88	2.53*
Men-women	−0.28**	0.92	2.31*
Young-old	−0.10	1.07	−1.49
Revised attrition model			
White-Black	−0.18*	0.96	0.83
White-Hispanic	−0.12	0.95	0.99
Men-women	−0.13	0.95	1.21
Young-old	−0.14	0.92	1.84

Note:  $N = 894$ . There were 401 Whites, 284 Blacks, 156 Hispanics, 524 men, 370 women, 710 below the age of 40 (young) and 184 above the age of 40 (old). \*\* $p < 0.01$ , \* $p < 0.05$ . The  $d$ 's reflect standardised mean differences in attrition scores. Negative values indicate the minority group has higher likelihood of termination. AIR = adverse impact ratio. Z is the test statistic for the difference in selection ratios. Percent variance explained for race (using all races as predictors within an ANOVA model) was 2.6% ( $p < 0.01$ ) for Full Attrition Scores, 1.6% ( $p < 0.01$ ) for Operational Attrition Scores, and 0.7% for Revised Attrition Scores ( $p = ns$ ).

Standardised mean differences between demographic groups on the continuous predictor variables can be seen in Table 4. According to Cohen's benchmarks, values of 0.2, 0.5 and 0.8 are small, moderate and large. Given predicted retention scores were higher for Whites, men and younger employees, variables with larger group differences across these demographics were considered. A rule of thumb was adopted that if the average standardised mean difference was 0.20 or greater (at least a small effect) and that variable was significantly related to turnover in favour of the majority groups, then that variable was removed from the analysis. Additionally, given the significant difference in decision rates against women, particular attention was given to variables adversely impacting women.

Based on this logic, sales commission had an average standardised mean difference that met the 0.20 cutoff and negatively impacted all minority groups and was thus removed. Units sold, while not adversely impacting minorities according to race or age, did adversely affect women ( $d = -0.27$ ,  $p < 0.01$ ), and this variable was also removed. Each of these variables were significantly related to turnover (Table 1). It should be noted that upon inspection, the relationships between predictors and turnover were linear, and variable importance weights from the random forest algorithm can be found in Figure 1 as well, which provides an indication of which variables contributed the most to turnover prediction while controlling for the effects of other variables.

The relationship between categorical variables with turnover and protected group status was examined next. Work location was significantly related to turnover (multiple  $R = 0.19$ ,  $p < 0.01$ ) and was significantly related to whether respondents were White (multiple  $R = 0.37$ ,  $p < 0.01$ ). It was also related to age (multiple  $R = 0.15$ ,  $p < 0.01$ ), though not with gender (multiple  $R = 0.03$ ,  $p = ns$ ). Given the effects with turnover and the significant

TABLE 4 Standardised mean differences by predictor

	White-Black	White-Hispanic	Men-Women	Young-Old
Job tenure	−0.11	−0.46**	0.19**	0.00
Department change	−0.07	−0.36**	0.12	−0.05
Internal candidate	−0.14	−0.01	0.13*	0.05
Education level	0.08	0.41**	0.05	−0.13
Supervisor change	−0.14	−0.48**	0.12	−0.01
Pay	0.17*	−0.21*	0.15*	−0.09
Pay change	0.06	−0.19*	0.11	−0.17
Sales commission	0.18*	0.18*	0.28**	0.30**
First call resolution	0.16*	−0.31**	0.08	0.14
Calls handled	−0.14	−0.10	0.13	0.00
Customer satisfaction	0.12	0.07	−0.03	−0.05
Units sold	−0.14	−0.23*	0.27**	−0.03
Sales efficiency	−0.23**	−0.33**	0.14*	0.09
Performance rating	0.17*	0.07	0.07	−0.08

Note:  $N = 894$ . \*\* $p < 0.01$ , \* $p < 0.05$ .

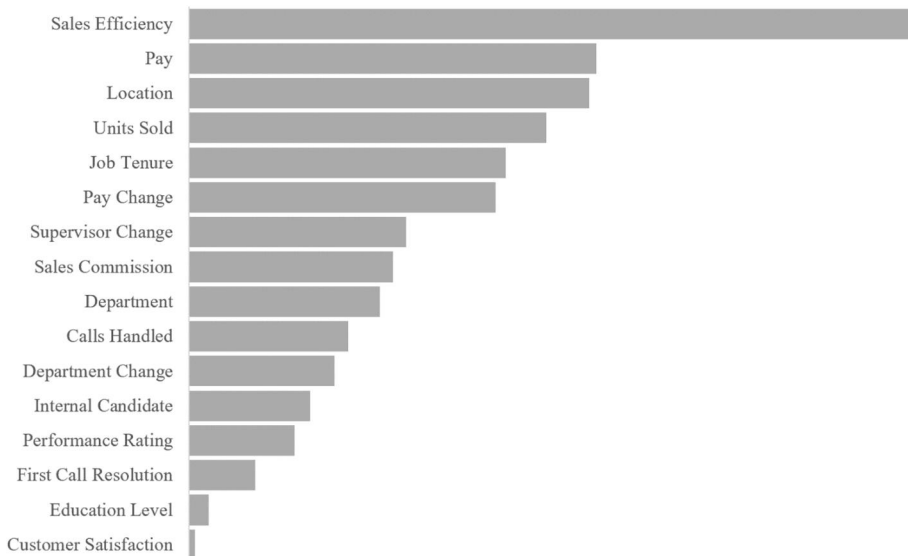


FIGURE 1 Variable importance weights for random forests operational attrition model. Note. Variable importance weights were based on random forest analysis using the entire sample ( $N = 894$ ). Values represent mean decrease in accuracy based on random permutations in the out-of-bag samples

group differences with race and age, the data were examined to determine whether locations with higher turnover also had more racial minorities and older employees. This was found to be the case for race, such that mean turnover rates by location covaried with mean percentage of Whites ( $r = -0.69$ ). No relationship was found when examining age. Given the effects in totally, location was removed from the model.

The second categorical variable was department. Department was not significantly or practically related to turnover (multiple  $R = 0.06$ ,  $p = \text{ns}$ ). It was related to whether respondents were White (multiple  $R = 0.35$ ,  $p < 0.01$ ) and age (multiple  $R = 0.09$ ,  $p < 0.05$ ), though not gender (multiple  $R = 0.07$ ,  $p = \text{ns}$ ). Additionally, there was no pattern of turnover and demographic membership by department. Given this, this variable was not removed in the model. Collectively then, three variables (sales commission, units sold, work location) were removed when forming the Revised Operational Attrition Scores.

After performing these removals and re-running the model, it was found that women were still adversely affected. Women still had significantly higher attrition probabilities than men ( $p < 0.05$ ), and although practical differences were small ( $d = 0.16$ ,  $\text{AIR} = 0.92$ ), the difference in selection rates remained significant ( $Z = 2.19$ ,  $p < 0.05$ ). There was not strong evidence of adverse impact against other groups. It should be noted that these newly created attrition scores were similarly related to turnover ( $r = 0.47$ ) as the original Operational Attrition Scores.

At this stage, some could make the argument that further revisions would not be needed, as gender differences were not practically meaningful, there was validity evidence and alternative procedures with slightly less adverse impact and similar validity had been examined. Despite this, an effort was made to sequentially remove variables until significant group differences were eliminated for gender, as additional alternative models might exist (without adverse impact but similar validity). The variable with the next largest difference between genders was employee tenure ( $d = -0.19$ ,  $p < 0.01$ ), which was significantly related to turnover ( $r = -0.11$ ). In addition to the prior edits, this variable was also removed, and random forests was run once again. The resulting scores, labelled Revised Operational Attrition Scores in Tables 2 and 3, were found to have no adverse impact. Effect sizes were small and non-significant for race (multiple  $R = 0.08$ ,  $p = \text{ns}$ ), gender ( $d = -0.13$ ,  $p = \text{ns}$ ) and age ( $r = 0.04$ ,  $p = \text{ns}$ ). Furthermore, AIRs were not violated, and the difference in decision ratios were also not significant (Table 3). Perhaps most importantly, the Revised Operational Attrition Scores maintained a healthy relationship with turnover ( $r = 0.46$ , Table 2). This is in comparison to the original Operational Attrition Scores ( $r = 0.48$ ), and the difference between these validity estimates was not significant ( $Z = 1.61$ ,  $p = \text{ns}$ ). Thus, with minimal decreases in model prediction, this procedure was able to fully remove adverse impact for the attrition scores.

## 4 | DISCUSSION

The output from attrition modelling and data-driven HR models can facilitate a number of initiatives. However, there has been little attention given to attrition modelling within the organisational sciences literature, and as a result there has not been great guidance on best practices when performing this work. To this effect, there has been little discussion of how attrition models and other data-driven HR models might facilitate employment decisions and therefore how such models might be subjected to employment laws. The current paper articulated the potential for group discrimination when performing attrition modelling. In doing this, I provided guidance on how to statistically test for model discrimination, connected attrition model discrimination to broader validity frameworks, and then demonstrated these principles using an actual attrition model.

Given the nature of this paper, the results will not be rehashed in detail. The approach initially suggested by Castille and Castille (2019) and more thoroughly articulated in this paper was successfully applied within this study and resulted in removal of adverse impact for this particular dataset. This paper, along with Speer et al. (2019), therefore serves as guides for practitioners in conducting attrition modelling. Too often in recent years, there has been a craze for new 'big data' methods and their potential in HR. While methods like ML offer immense promise for improving the prediction of work outcomes, implementation in practice is complex, with sometimes unintended consequences. This paper not only provides an applied example of how an attrition model based on ML is formed, but it provides a discussion and then empirical demonstration of how group differences in attrition scores must be considered.

If attrition models or other data-driven HR models do influence employment decisions, then prior to implementation of such models researchers are advised to (1) develop algorithms without any protected class variables and collect evidence of model validity, (2) test the developed scores for adverse impact and then (3) if adverse impact is found, attempt to remove adverse impact by removing variables or altering variable weights (e.g. via Pareto optimisation or ML-based procedures). If model revisions result in reduced adverse impact while maintaining model validity, such revisions should be accepted for future use. Furthermore, if a model has adverse impact but an equally valid alternative with less adverse impact cannot be found, the model should still be legally viable. This might be relevant, as suggested by a reviewer, if the original operational model has substantial adverse impact which cannot be reasonably removed while maintaining similar levels of validity. Despite this, users may still opt for caution in such situations. This is because employment decisions resulting from models such as those producing attrition scores should serve an important business purpose, and while there is precedence for alternative procedure considerations in employment litigation, none yet exists for attrition scores. Thus, caution is likely wise. Additionally, attrition scores resulting in adverse impact may harm organisational diversity goals, and thus users would need to strongly consider whether the gains from the model exceed the costs incurred by group differences. Thus, researchers should always seek to maximise validity while simultaneously reducing adverse impact.

As alluded to early in this paper, how attrition scores affect diversity goals and adverse impact is quite complicated. If those with low attrition risk receive favourable employment decisions (e.g. to maximise return-on-investment given the limited number of available rewards), the majority group would be coded differently than if those with high attrition scores received favourable employment decisions (e.g. rewards given to prevent those most at risk from leaving). Ultimately, how companies use attrition scores will differ by context, and the intent of this paper was addressing adverse impact and validity issues rather than expanding upon the optimal usage of attrition models. Nonetheless, it is a fascinating issue and makes attrition modelling complex as it pertains to adverse impact, as opposed to more traditional HR practices such as employee selection. It should also make companies thoroughly consider whether using attrition scores would harm or hurt diversity goals. Given the possible negative repercussions of using attrition scores for individual decisions, it is likely prudent to use attrition modelling for understanding group trends and only using for individual decisions if there is a highly business relevant reason to do so.

## 4.1 | Limitations and areas for future research

The exact attrition model that was tested in this study was not intended to generalise. Rather, the method by which researchers address adverse impact and validity should, such that this paper provides applied researchers a step-by-step framework when confronting attrition modelling. Despite this, researchers and practitioners are advised to investigate nuances to these recommendations. For example, one weakness of this study was that a relatively small dataset was used. The sample size has obvious implications on adverse impact significance testing, such that the likelihood for significant group differences increase as sample size grows. Like with employment litigation, it makes sense to pair practical significance with statistical significance (Murphy & Jacobs, 2012), else any minuscule difference in scores will result in statistically significant effects. This may be particularly important when revising attrition models to reduce adverse impact, as fluctuations in sample size impact inferential statistics. As an example, model edits might be conducted on a calibration sample to remove adverse impact only to have adverse impact later occur in operational use because the sample size is larger. Once again, this speaks to the importance of considering both practical and statistical significance. Also relevant to sample size is determining whether employee groups should be pooled together; traditional adverse impact analyses will separate calculations into groups of employees who are similarly situated in the organisation (Morris et al., 2019), and it might make sense to segment the sample by job or other differentiating feature within attrition models too. The decision might ultimately be



based on how models are used (e.g. are interventions or decisions provided consistently across all members of the company, irrespective of job).

The number of variables used in this study was also small. Increases in the number of variables should lead to improved model performance (i.e. higher validity), but it does become increasingly difficult to perform sequential variable removal in efforts to reduce adverse impact, as was done in this study. This is especially the case for complex ML algorithms with many variables and when trying to minimise adverse impact for multiple group characteristics. The adverse impact removal strategy in this study was univariate and therefore simplistic. Organisational researchers are encouraged to explore more advanced optimisation procedures that weight variables in a way that balances algorithm validity and adverse impact (Wadsworth et al., 2018), and when doing so consider whether such procedures adhere to existing laws.

Finally, this study used a local validation approach, but there might be need for non-local validation of attrition models. Validation is straightforward when attrition models are custom built within each organisation. However, off-the-shelf attrition models from external testing vendors and for HR software vendors that offer attrition dashboards do exist. In such cases, criterion-related validity could be locally established by showing those models predict turnover in the new organisation prior to usage of those models. However, one wonders whether non-local validity evidence can be applied to these organisations as well. For example, could the validity of an attrition model be 'transported' to a new organisation if the contexts are deemed similar enough? To my knowledge, issues of transportability have only been applied to selection contexts (Principles, 2018). Future work should consider how similarity would be established for transportation of other types of data-driven HR models. Would transportation occur at the job level? Would it be based on company similarity? What tools would be used to infer this? These are questions without clear answers, but which have applied implications.

## 4.2 | Conclusions

People analytics and its related insights have come at the forefront of organisational research in recent years. Practices such as attrition modelling can provide value to organisations. However, data-driven HR algorithms should be considered in line with traditional adverse impact and validity frameworks. The current paper explored these considerations in detail and then demonstrated how to test and revise attrition models to increase legal justification for use. As the frequency of this and related practices increases, it is vital for scholars and practitioners to consider how new methods are implemented. The current paper does this by connecting disparate treatment, adverse impact and validity theories to attrition modelling. Researchers developing attrition model should consider these factors when predicting turnover.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

## ORCID

Andrew B. Speer  <https://orcid.org/0000-0002-3376-2103>

## ENDNOTES

- <sup>1</sup> Note that the intent of this paper is not to compare competing ML methods, as modern prediction methods are already well established as effective predictors in work settings (Putka et al., 2018). Rather, within this study, ML is simply used as a tool to facilitate the creation of attrition scores.
- <sup>2</sup> Additional methods such as qualitative reviews and deeper inquiry to more fully understand turnover phenomena are important and should also be conducted.
- <sup>3</sup> This paper is largely framed within the US legal context given greater familiarity in this domain, though many of the principles will be applicable in other legal environments.

- <sup>4</sup> Technically, the 'standard deviation' analysis does not express differences in standard deviations but rather standard errors, which are a function of the standard deviation and sample size.
- <sup>5</sup> Although validity insulates a test when it has AI, most organisations would prefer to avoid any prima facie case from arising altogether.
- <sup>6</sup> Optimal settings for the neural network were a softmax layer stacked onto a layer of five hidden units, stacked onto a layer of 20 hidden units, 25 epochs with a batch size of 32, with dropout of 0.1 between the hidden layers and with RELU activations used in the lower layers.
- <sup>7</sup> With a very large number of variables, I often engage in variable reduction prior to random forests or other ML algorithms, as I have found this often results in model performance gains. However, with so few predictors in this study, this step was not necessary.
- <sup>8</sup> Adverse impact results for involuntary turnover and voluntary turnover were not presented in Table 3 because (1) group predictions did not differ across turnover types (i.e. standardised mean differences were not meaningfully different when comparing overall turnover, voluntary turnover and involuntary turnover predictions), (2) because the intent was on predicting all forms of turnover simultaneously and therefore using the overall turnover likelihood scores and (3) for parsimony.

## REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*.
- Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal*, 26, 1–11.
- Becker, W. J., & Cropanzano, R. (2011). Dynamic aspects of voluntary turnover: An integrated approach to curvilinearity in the performance-turnover relationship. *Journal of Applied Psychology*, 96, 233–246.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Castille, C. M., & Castille, A.-M. R. (2019). Disparate treatment and adverse impact in applied attrition modeling. *Industrial and Organizational Psychology*, 12, 310–313.
- Cohen, D., Tison, E., & Fortney, D. S. (2019). Structuring a traditional EEO adverse impact analysis. In S. B. Morris, & E. M. Dunleavy (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 49–70). Routledge.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. Erlbaum.).
- Conway, J., & Frick, S. (2017). A predictive turnover model for global private banking relationship managers. Paper presented at the 32nd Annual Conference of the Society for Industrial & Organizational Psychology, Orlando, FL.
- De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, 92, 1380–1393.
- Deere, D. R., & Pearce, J. E. (2019). Analyzing reductions in force and other termination decisions. In Morris, S. B., & Dunleavy, E. M. (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 239–257). Routledge.
- Doverspike, D., Arthur, W. Jr., & Flores, C. (2019). Analyzing EEO disparities in pay: A primer on structuring analyses. In Morris, S. B., & Dunleavy, E. M. (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 197–215). Routledge.
- Eissenstat, K. (2016). Lies, damned lies, and statistics: The case to require "practical significance" to establish a prima facie case of disparate impact discrimination. *Oklahoma Law Review*, 68, 641–675.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, 43, 38290–39315.
- Es-Sabahi, N., & Deluca, D. (2017). Utilizing machine learning to predict turnover. Paper presented at the 32nd Annual Conference of the Society for Industrial & Organizational Psychology, Orlando, FL.
- Gibson, C., Koenig, N., Griffith, J., & Hardy, J. H. (2019). Selecting for retention: Understanding turnover prehire. *Industrial and Organizational Psychology*, 12, 338–341.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Hom, P. W., Mitchell, T. R., Lee, T. W., & Griffeth, R. W. (2012). Reviewing employee turnover: Focusing on proximal withdrawal states and an expanded criterion. *Psychological Bulletin*, 138, 831–858.
- International Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977).
- King, A. G., & Mrkonich, M. (2016). "Big Data" and the risk of employment discrimination. *Oklahoma Law Review*, 68, 555–584.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Lewis, N. (2019, November). AI-related lawsuits are coming. Retrieved from <https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/ai-lawsuits-are-coming.aspx>

- Lundquist, K. K., Locklear, T. S., & Lippstreu, M. (2019). Using your data wisely: Proactive monitoring of employment disparities. In S. B. Morris, & E. M. Dunleavy (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 261–277). Routledge.
- Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR analytics. *International Journal of Human Resource Management*, 28, 3–26.
- Martin, D. C., Bartol, K. M., & Kehoe, P. E. (2000). The legal ramifications of performance appraisal: The growing significance. *Public Personnel Management*, 29, 379–406.
- McCloy, R. A., Smith, E. A., & Anderson, M. G. (2016). Predicting voluntary turnover from engagement data. Paper presented at the 31st Annual Conference of the Society for Industrial & Organizational Psychology, Anaheim, CA.
- Morris, S. B., & Dunleavy, E. M. (2019). *Adverse impact analysis: Understanding data, statistics, and risk*. Routledge.
- Morris, S. B., Dunleavy, E. M., & Lee, M. (2019). Many 2x2 tables: Understanding multiple events in adverse impact analyses. In Morris, S. B., & Dunleavy, E. M. (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 147–168). Routledge.
- Murphy, K. R., & Jacobs, R. R. (2012). Using effect size measures to reform the determination of adverse impact in equal employment litigation. *Psychology, Public Policy, and Law*, 18, 477–499.
- Obenauer, W. G. (2019). Are all voluntary attritions created equally? Understanding the need to incorporate employee diversity into attrition modeling. *Industrial and Organizational Psychology*, 12, 302–305.
- Oswald, F. L., Dunleavy, E. M., & Shaw, A. (2019). Measuring practical significance in adverse impact analysis. In Morris, S. B., & Dunleavy, E. M. (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 72–92). Routledge.
- Principles for the Validation and Use of Personnel Selection Procedures. (2018). *Industrial and organizational psychology: Perspectives on science and practice*, 11, 2–97.
- Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods. *Organizational Research Methods*, 21, 689–732.
- R Core Development Team. (2007). *R: A language and environment for statistical computing*. Foundation for Statistical Computing.
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, 29, 615–620.
- Rosett, C. R., & Leinweber, K. (2017). Predicting frontline turnover: A practical approach yielding early results. Paper presented at the 32nd Annual Conference of the Society for Industrial & Organizational Psychology, Orlando, FL.
- Rubenstein, A. L., Eberly, M. B., Lee, T. W., & Mitchell, T. R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, 71, 23–65.
- Siskin, B., & Schmidt, N. (2019). Proper methods for statistical analysis of promotions. In S. B. Morris, & E. M. Dunleavy (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 171–196). Routledge.
- Speer, A. B., Dutta, S., Chen, M., & Trussell, G. (2019). Here to stay or go? Connecting turnover research to applied attrition modeling. *Industrial and Organizational Psychology*, 12, 277–301.
- Strickland, W. J. (Ed.). (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions*. U.S. Army Research Institute for the Behavioral and Social Sciences. Retrieved from <https://apps.dtic.mil/dtic/tr/fulltext/u2/a440522.pdf>. (Technical Report 1172).
- Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: An application to recidivism prediction. *ArXiv*, abs/1807.00199, 1–5.
- Werner, J. M., & Bolino, M. C. (1997). Explaining U.S. courts of appeals decisions involving performance appraisal: Accuracy, fairness, and validation. *Personnel Psychology*, 50, 1–24.
- Zhou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67, 301–320.

**How to cite this article:** Speer AB. Empirical attrition modelling and discrimination: Balancing validity and group differences. *Hum Resour Manag J*. 2021;1–19. <https://doi.org/10.1111/1748-8583.12355>