Social Policy and Society (2023), page 1 of 14
doi:10.1017/S1474746423000428

ARTICLE

(in) Accuracy in Algorithmic Profiling of the Unemployed – An Exploratory Review of Reporting Standards

Patrick Gallagher and Ray Griffin

Department of Management and Organisation, South East Technological University, Ireland
Corresponding author: Ray Griffin; Email: ray.griffin@setu.ie

(Received 20 April 2022; revised 14 March 2023; accepted 20 March 2023)

Public Employment Services (PES) increasingly use automated statistical profiling algorithms (ASPAs) to ration expensive active labour market policy (ALMP) interventions to those they predict at risk of becoming long-term unemployed (LTU). Strikingly, despite the critical role played by ASPAs in the operation of public policy, we know very little about how the technology works, particularly how accurate predictions from ASPAs are. As a vital first step in assessing the operational effectiveness and social impact of ASPAs, we review the method of reporting accuracy. We demonstrate that the current method of reporting a single measure for accuracy (usually a percentage) inflates the capabilities of the technology in a peculiar way. ASPAs tend towards high false positive rates, and so falsely identify those who prove to be frictionally unemployed as likely to be LTU. This has important implications for the effectiveness of spending on ALMPs.

Keywords: Public employment services; labour market profiling; automated statistical profiling algorithms; active labour market policy

Introduction

Public Employment Services (PES) increasingly use automated statistical profiling algorithms (ASPAs) to predict unemployed jobseekers that are at risk of becoming long-term unemployed (LTU). Categorising jobseekers allows PES to target more intensive activation measures (McDonald *et al.*, 2003; Desiere *et al.*, 2019) at those most in need of support. A key policy objective of PES is to reduce LTUs given their associated economic scarring and sociological well-being (Brandt & Hank, 2014) effects and long-term impacts on the labour market (Strandh & Nordlund, 2008) and social cohesion. However, ALMPs involve expensive, intensive, human services of personalised mentoring and counselling (Senghaas *et al.*, 2019), training and sheltered employment supports; such policies can cost as much as 0.6 percent of gross domestic product (GDP) in some countries (Pignatti & Van Belle, 2018). Therefore, algorithms hold the potential to lower costs and improve service impacts by rationing access to expensive ALMPs in ways that reduce deadweight loss (Loxha & Morgandi, 2014; O'Connell *et al.*, 2009) – reducing spending on those who do not need ALMPs. However, the potential of ASPAs to increase efficiency and lower costs rests on their ability to accurately identify those most at risk of LTU (Desiere & Struyven, 2021).

Across the Organisation for Economic Co-operation and Development (OECD), thirteen countries currently employ ASPAs to predict those most at risk of LTU to target expensive ALMPs. Each of the thirteen countries that employ ASPAs uses entirely different systems – statistically and administratively; no two are alike. Reported accuracy rates for the ASPAs operating in the OECD range from 60 per cent to 86 per cent; however, it is impossible to interpret

these rates because they are not accompanied by a methods statement that outlines what the rate means in terms of classification and misclassification (Desiere *et al.*, 2019). In medical science, where the use of predictive modelling is well established, it is best practice when testing and reporting on the accuracy of predictive models to provide multiple measures for accuracy that relate directly to the intended use of the model (Reilly & Evans 2006; Ferrante di Ruffano *et al.*, 2012; Kappen & Peelen, 2016). Therefore, as a vital first step in assessing the operational effectiveness and social impact of ASPAs the current paper reviews the method of reporting accuracy as a single measure.

The analysis in this article builds on a project funded by European Commission to develop a next-generation PES tool. As part of that project, we evaluated existing ASPAs to identify why their adoption had not achieved a dominant design. A core feature of the assessment was to conduct a review of the reported accuracy rates of the thirteen first-generation ASAPs operating in the OECD. For data, we assemble the reported accuracy rates supplemented with associated methodological materials and empirical research on the accuracy of ASPAs. Drawing on methodological insight from medical and data science, two fields of scholarly inquiry that both have long-established histories of testing predictive models and algorithms, we review the method of reporting accuracy rates for the ASPAs currently operating in the OECD. In our review, we find the current method of reporting to be opaque and lacking in methodological rigour. In particular, reporting accuracy using as a single measure inflates the capabilities of the technology and is meaningless when trying to assess the operational efficiency of ASPAs and therefore comes across as a political gesture to appear accurate. In particular, based on our review of empirical testing of ASPAs, we demonstrate that providing a single measure for accuracy obfuscates many false-positive errors in which the frictionally unemployed are mistakenly classified as long-term unemployed. Moreover, given that the justification for the use of ASPAs hinges on their ability to predict LTU accurately and the high number of errors identified in the course of our review, we question the feasibility of the technology (O'Connell *et al.*, 2009 Matty 2013; Desiere *et al.*, 2019). Finally, in assessing the social impact of ASPAs we argue that what is at stake here is not just identifying a malfunctioning technology or costly ineffective PES; instead, it is the possibility that a technology designed to assist the unemployed are pushing them further into marginalisation (Loopstra & Tarasuk, 2015; Sage, 2015).

The call for algorithmic profiling to ration PES councillor workload

Access to unemployment insurance or benefits has always been algorithmic – a finite sequence of well-defined instructions or tests to resolve if an applicant is entitled to social transfers and support services. However, unemployment has a settled and practically operationalised International Labour Organisation (ILO) definition since 1954 of not working, being available for work and seeking work (Demazière, 2014; Boland and Griffin, 2015). This ILO definition has proved remarkably potent amongst the 175 affiliated countries. Indeed, this precise administrable construct has colonised how most states address poverty, operationalised as a simple, reductive algorithmic test to access PES social transfers and support services. Traditionally application forms were used to elicit information for the PES to test against defined access rules such as citizenship and social insurance contributions, with bureaucrats using discretion circumscribed by administrative oversight. However, since the 1970s increased focus has been placed on the 'seeking work' element of the ILO definition, a development often described as the 'activation turn' (Bonoli, 2010) when welfare dependency, persistent poverty, and labour market exclusion of LTUs, became a significant policy concern.

Central to the 'activation turn' is the increasing use of regular face-to-face meetings with PES caseworkers to address labour market reintegration through fine-grained targeted supports. However, much research demonstrates that ALMPs are expensive and time-consuming and have placed added pressure on increasingly stretched caseworkers (Senghaas *et al.*, 2019; Kaufman, 2020).

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

Indeed, the most authoritative study, a meta-analysis of 200 evaluations of specific ALMPs found that in the short run average impacts were close to zero (Card *et al.*, 2018). It may well be the case, that ALMPs are targeted at the wrong people. Furthermore, the limits experienced by PES in the implementation of ALMPs have contributed to the growing emphasis amongst national and international commentators on maintaining a low ratio of front-line PES caseworkers/counsellors to service users, with the average within the European Union being around 1:150, against an ILO recommendation of 1:100. Therefore, with the unemployment rate and thus the number of unemployed highly variable, particularly in the context of the global financial crisis, pandemic, and intermittent periods of economic growth – rationing access to PES is a priority. In the context of lack of capacity, PES have the unpalatable choice of reducing their service mix to a level where it can offer universal coverage; a first-come, first-served principle, queueing, or some form of differentiation (Hasluck, 2008) such as profiling.

Profiling has a range of benefits for PES, increasing efficiency whilst lowering costs and increasing the capacity to tailor services to individual needs (Desiere *et al.*, 2019), aligning with a key objective of identifying those most distant from the labour market and allocating scarce resources to their most vulnerable jobseekers. The literature distinguishes three types of profiling: rule-based, case-worker-based, and statistical profiling (Loxha & Morgandi, 2014). Rule-based profiling uses eligibility criteria established through administrative data to determine the required level of support. Caseworker profiling relies on the human judgement and experience of the PES worker to profile individuals. Statistical profiling uses a statistical model related to client characteristics (mainly hard characteristics such as gender, age, education, occupation, and work experience, but sometimes soft characteristics such as motivation and social networks). Statistical profiling is seen as more accurate than simple rule-based systems or human decision-making (Zejnilović *et al.*, 2020; Desiere & Struyven, 2021) and is often considered an objective way of segmenting jobseekers and prioritising the rationing of scarce resources (Martin & Grubb, 2001).

ASPAS in practice

The USA was the first significant deployment of ASPAs, with the Unemployment Compensation Amendment of 1993 mandating States to develop profiling systems. The Worker Profiling and Re-employment Services (WPRS) system aspires to predict the probability of exhausting unemployment insurance benefits by each new claimant. In 1994 Australia (McDonald *et al.*, 2003; Lopez, 2019) developed a different approach to profiling, continually revising and refining the Job Seeker Classification Instrument (JSCI) to identify the risk of LTU; caseworkers use this as input to deciding on supports. Denmark and Germany followed in the mid-2000s. Many more countries have since implemented systems or experimented with them (Hasluck, 2008). At the time of writing, high-profile examples include segmenting tools such as The Netherlands – Work Profiler (Wijnhoven & Havinga, 2014), Croatia – Statistically Assisted Profiling StAP (Bejaković & Mrnjavac 2018), Sweden – Assessment Support Tool AST (Loxha & Morgandi, 2014), Irish Probability of Exit PEX (O'Connell *et al.*, 2009; McGuinness *et al.*, 2014), Finland – Statistical Profiling Tool (Riipinen, 2011), and others with more a case management approach Denmark – Job Barometer (Rosholm *et al.*, 2004; Larsen & Jonsson, 2011), Poland (Niklas *et al.*, 2015). The United Kingdom experimented with a profiling model but decided not to implement it as a practical instrument following concerns about the model's accuracy (Matty, 2013), many other implementations are being reformed or abandoned.

The OECD, alongside other international organisations, and European PES play a crucial role in the oversight of algorithmic profiling of the unemployed. Several reports and cross-country comparisons suggest that technology is an efficient and cost-cutting means of delivering services to the unemployed with the added benefit of increased personalisation (Loxha & Morgandi, 2014; Desiere *et al.*, 2019). Furthermore, while acknowledging some concerns with ethics and transparency, key reports suggest that ASPAs have been successfully deployed across a range of

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

European PES (Desiere *et al*., 2019) and predict the likelihood of LTU with a high degree of accuracy-reported accuracy rates to a range between 60 per cent and 86 per cent. These key reports and their source material from which they draw accuracy rates, do not offer any method statement indicating how the accuracy rate was composed. In particular, none offers a breakdown of the false positive rate and false negative rate, over what timeframe accuracy was determined, the sample used, sampling uncertainty and other typical details that would reveal a rigorous consideration of the accuracy of the ASPA was undertaken. Indeed, to a statistician or data scientist, this opacity suggests reported accuracy is more 'impression management' (Goffman, 1959), more an effort to explain away negative outcomes, than a genuine attempt to offer PES administrator information about the accuracy of a tool they might rely upon.

While American and Australian ASPA deployments are automated decision-making, most European deployments are described as having a human override, or being decision support systems. Research on understanding this form of hybridic human-algorithm decision-making is in its infancy (Bader & Kaiser 2019) and it is unclear to what extent humans rely upon or defer to machines (Danaher 2016). For PES caseworkers and administrators to calibrate their own reliance on ASPAs (Lipsky 2010), they need to understand the limits of affordances of this technology. We are now only getting to grips with the importance of reporting accuracy, so for example IZA developed an evidence map (evidence map for statistical profiling of unemployed jobseekers (iza.org)) that outlines the available methodological and secondary data for the various implementations of statistical profiling operating in the OECD (Van Landeghem *et al*., 2021). The map indicates the paucity of information on the accuracy of ASPAs, with oversight for five countries – Ireland, Denmark, Latvia, Norway, and Sweden – provided by just one cross-country report (Desiere *et al*., 2019).

So, over the past twenty-five years of experience with ASPAs, one-third of OECD countries, thirteen of the thirty-six members, use algorithmic profiling to deliver public employment services (PES), by and large, to categorise the unemployed into two groups – high and low risk of long-term unemployment, to target expensive ALMPs. While not yet incumbent or standardised across the OECD based on perceived benefits to efficiency and cost-effectiveness, the technology has grown popular. However, the promise of increased efficiency and personalisation of service delivery that ASPAs hold hinges on their ability to decipher those most distant from the labour market from the frictionally unemployed with a sufficient degree of accuracy. Unfortunately, based on the current approach to reporting accuracy, it is not possible to establish how accurately ASPAs identify the risk of LTU – in particular, it is unclear how providing a single rate for accuracy allows the reader to establish the level of classification and misclassification.

A research agenda for ASPAs

Across the OECD, thirteen countries currently employ ASPAs to parse those at risk of LTU from the frictionally unemployed. Because this instance of digitisation of economy and society is the state-in-action at the point of care to some of its most vulnerable citizens, we can see a premonition of the troubles a digital future might bring - the translation of bureaucratic logics into an algorithmic authority (Lustig & Nardi, 2015; Griffin *et al*. 2020). Algorithmic authority refers to the power of algorithms to manage human action and influence what information is accessible to users – cases include algorithmically curating the news and social media feeds, evaluating work performance, matching on dating sites, and hiring and firing employees. In the case of ASPAs the technology wields algorithmic authority in determining the type of PES (active or passive) offered to the unemployed. The shift to algorithmic authority in PES has the potential to be problematic if risks and disadvantages are not well understood.

Algorithms that categorise individuals are vulnerable to two types of classic mistakes false positives and false negatives. First, false positives are the incorrect assignment of a label, such as labelling someone a terrorist when they are not. Alternatively, false negatives incorrectly exclude

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

someone from a category, for example, identifying someone as not a terrorist when they are. The susceptibility of algorithms to two types of errors (false positive and false negative) in the context of what PES hope to achieve by applying the algorithm raises important questions about how a single measure of accuracy can render any insight into the operational effectiveness or social impact of ASPAs. For example, the technology aims to increase efficiency by identifying and targeting those most at risk of LTU. It follows that to establish operational efficiency, we need multiple accuracy measures for both the LTU group and the frictional group, including how many LTU people are correctly and incorrectly classified and how many frictionally unemployed are correctly and incorrectly classified. In particular, we need to know each group's error rate – how many frictionally unemployed are mistakenly classified as high risk (false-positive) and how many of the LTUs are mistakenly classified as low risk (false-negative). Furthermore, it is unclear why accuracy is not reported clearly and concisely to allow access to performance data to scholars from the social sciences that are best equipped to assess the social impact of the technology.

As a result, we know little about the impact of errors in statistical profiling on the lives of the unemployed. For example, errors within ASPAs are performative, dictating with algorithmic authority (Lustig & Nardi 2015) the type of active labour market policy offered to individuals. Therefore, those mistakenly classified as high risk will potentially be unnecessarily placed in active labour market programmes and may be subject to higher levels of welfare conditionality associated with these programmes. Alternatively, those mistakenly predicted as low-risk of LTU would be denied access to a suite of supports – two areas that are crucial to assess operational effectiveness and social impact. Therefore, what is at stake here is not simply identifying a technology that is not functioning as envisaged or an ineffective, costly PES, but the possibility that a technology designed to assist the unemployed by increasing efficiency and personalisation of service delivery is pushing them further into marginalisation.

Methodology

The analysis in this article builds on a project funded by European Commission to develop a next-generation PES tool. As part of the EU Horizon 2020 funded HECAT project, we evaluated existing ASPAs to identify why their adoption had not achieved a dominant design. A core feature of the assessment was to review the reported accuracy rates of the thirteen first-generation ASAPs. For data, we assemble reported accuracy rates for the ASPAs operating across the OECD, supplemented with the associated reporting materials around oversight and empirical testing of the technology.

Policy evaluation applies assessment principles and methods to appraise the content, implementation or impact of a policy or programme. It is an analytical method to gauge a policy's merit, worth, and utility. Wollmann (2003) suggests that policy evaluation should achieve two things when applied as an analytical tool. Firstly, all the information pertinent to the assessment of the performance of the policy or program should be gathered, and secondly, this information should be fed back into the policymaking process. We could see that assessing the technology's 'actual accuracy rate' was an essential first step in evaluating the merit, worth, and social impact of ASPAs. To this end, we did two things: firstly, we assembled the reported accuracy rates for ASPAs operating in the OECD based on a systematic literature review (Sundberg, 2017), $n = 146$, and combined these with the associated reporting, audit, empirical tests, and methodological materials. In evaluating the assembled material, it became clear that interpreting the reported accuracy rates, in particular establishing the level of classification and misclassification, contained technical challenges that could potentially hamper our efforts to feed the findings of the review back into the policymaking process.

From a technical perspective, ASPAs posed two specific methodological challenges that made it challenging to establish the operational efficiency of the technology. Firstly, the mathematical

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

complexity of the ASPAs, and secondly, the opaqueness of the data science terminology, both of which make the determinaton of accuracy abtruce. To overcome these challenges, we drew on methodological insight from two fields with long histories of testing and appraising predictive models – medical (Swets, 1988; Reilly & Evans, 2006; Kappen & Peelen, 2016; Kappen *et al.*, 2018) and data science (Tashman, 2000), and read the literature on ASPAs against this grain. Drawing on these resources allows us to understand the issues around reporting standards for establishing the effectiveness of predictive algorithms, in particular, false positives and negatives and cut-offs (Allhutter *et al.*, 2020) in the thirteen ASPA deployments we reviewed. An important theme running through these evaluations is the need to carefully assess what metrics are used to assess effectiveness (Ferrante di Ruffano *et al.*, 2012) and warnings about the capacity for various metrics to give a distorted picture of predictive models' ability to improve clinical outcomes (Reilly & Evans, 2006). This approach allowed us to decipher the mathematical complexity and brought clarity to the opaqueness created by data science terminology used in the current method of reporting accuracy.

Reviewing reported accuracy rates

In the following section, we assess the metrics used to report the accuracy of ASPAs (Ferrante di Ruffano *et al.*, 2012) and demonstrate the capacity for various metrics to give an inflated sense of accuracy and thus create a distorted picture of operational efficiency (Reilly & Evans, 2006). In particular, we show how providing a single measure for accuracy obfuscates the high numbers of frictionally unemployed misclassified as high risk (false positives) and communicates an inflated sense of ASPAs capability to predict the likelihood of LTU.

Table 1 is revised and extended from a recent publication by the OECD, which reviews current iterations of ASPAs and assembles the accuracy rates reported by individual countries (Georges, 2008; Desiere *et al.*, 2019). It is unclear if the measure is standardised and thus comparable across the various case countries in the reported accuracy rates. Such data would typically be reported with a method statement and a more significant effort to explain to non-expert users at the point of reporting what this accuracy measure means. Without this, the measure is meaningless and comes over as a gesture to appear accurate whilst being obtuse about actual accuracy.

Additionally, it is unclear how the accuracy of predictive modelling, which generates four result categories, can be meaningfully communicated to the reader through a single rate. By way of explanation, Figure 1 is a two-by-two contingency table demonstrating all four possible results generated by predictive modelling. The figure splits vertically to show the two distinct groups within our sample, the LTU and the frictionally unemployed, and horizontally to delineate the high and low-risk groups predicted by the ASPA. Both risk groups have two measures for accuracy – one for correct and another for incorrect predictions. To establish how accurately ASPAs predict the likelihood of LTU, we must consider each of these measures separately. Therefore, it is unclear how accuracy can be meaningfully reported using a single measure.

In the course of our efforts to uncover the meaning of the single accuracy rate reported by the OECD in terms of classification and misclassification, we drew on materials relating to oversight and empirical testing of ASPAs. In the little empirical research available in the public realm, primarily on the British (Matty, 2013), Danish (Rosholm *et al.*, 2004), Belgian (Desiere & Struyven, 2021), and Austrian tools (Allhutter *et al.,* 2020), it appears that what is being reported by the OECD as the overall accuracy is a measure developed by data scientists called 'forecast accuracy' (Swets, 1988). This measure focuses solely on the target group (LTUs in the case of ASPAs) and is explicitly constructed to ignore misclassifications in the non-target group (Swets, 1988). In operational terms, forecasting accuracy only picks up how many LTU people were missed (false negatives), and ignores how many frictionally unemployed people were mislabelled at LTU (false positives).

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

Table 1. Adapted and extended by the authors from Desiere *et al*. (2019) and Georges (2008)

Country	Self-reported efficacy	Purpose	Statistical method	Source
Ireland- PEX	50–69%	Identifying those at risk of LTU (6 mths)	Probit regression	O'Connell et al. (2009), Griffin et al. (2020)
Austria- AMAS	80–85%	Identifying those at risk of S/M/L TU (3/7/24 mths)	Logistic regression	Desiere et al. (2019)
Denmark- Job Barometer	66%	Identifying those at risk of LTU (6 mths)	Logistic regression	Rosholm et al. (2004); Larsen & Jonsson (2011)
France- Intelligence Emploi	70–80%	Identifying those at risk of LTU (6 months)	Neural network	OWALGROUP (2019)
Australia- JSCI	Not reported	Identifying those at risk of LTU (12 mths)	Logistic regression	Ponomareva & Sheen (2013), Lipp (2005)
Croatia- StAP	69%	Identifying those at risk of LTU (12 mths)	Logistic regression	Botrić (2017)
				Flesicher (2016)
Finland- Risk Profiling Tool	89%	Identifying those at risk of LTU (12 mths)	Logistic regression	Riipinen (2011); Behncke et al. (2007)
Belgium- VDAB	67%	Identifying those at risk of LTU (6 mths)	Random forest	Desiere et al. (2019)
Italy	70–90%	Identifying those at risk of LTU (12 mths)	Logistic regression	OECD (2019a)
Latvia	60–70%	Identifying those at risk of LTU (12 mths)	Factor analysis	Desiere et al. (2019)
				OECD (2019b)
Netherlands- WorkProfiler	70%	Identifying those at risk of LTU (12 mths)	Logistic regression	Wijnhoven and Havinga (2014), Hasluck (2008)
New Zealand- SEM	63–83%	Identifying those at risk of LTU (6 mths)	Random Forrest and Gradient boosting	Ministry for Social Development (2018)
Sweden-AST	85–90%	Identifying those at risk of LTU (6 mths)	Logistic regression	Loxha & Morgandi (2014)

2x2 Contingency Table

	Risk Scores ↓		***LTU Group***	***Frictional Group***
	100	High Risk Group	True Positive	False Positive
Cut Off	. . . 1	Low Risk Group	False Negative	True Negative

Figure 1. Two-by-two contingency table for predictive modelling.

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

Beyond this precise meaning of 'forecast accuracy' (Tashman, 2000), usefully the modifier 'forecast', moderates the perception of certainty conveyed by the measure. However, in reporting the accuracy rates of ASPAs the modifier is omitted, lending the technology an inflated sense of capability, hinting that the reported accuracy might also include false positives.

Given the lack of method statements around the reported accuracy rates outlined in Table 1, it is possible, although unlikely, that the accuracy rates are a composite of both the false negatives and the false positives.

Reporting accuracy as a single measure is misleading because it ignores errors, false positive and false negative results, meaning neither the frictionally unemployed who are mistakenly classified as high risk, nor the LTU who are mistakenly classified as low risk, are unaccounted for in the reported rate. This approach to reporting has significant implications for the operational viability of ASPAs. For example, ASPA can generate a high number of true positives while simultaneously generating a high number of false positives. If we consider this possibility in the light of the intended usage of ASPAs – separating those most at risk of LTU from the frictionally unemployed – the technology would be ineffective because the high-risk group (who receive extra support) would contain many frictionally unemployed who do not require support. In our next section, we demonstrate that providing multiple measures for accuracy (including error rates) renders a more informative picture of ASPAs' ability to predict LTU and lends crucial insight into the operational (in) effectiveness of ASPAs. In particular, we show that the false-positive rate obfuscated under the current reporting system is a crucial performance metric for assessing the operational effectiveness of ASPAs.

A new approach to reporting accuracy

In the following section, we demonstrate a method of reporting accuracy (including error rates) that provides a clear, concise, and accessible account of how (in) accurately ASPAs predict the likelihood of LTU. The purpose of this is to raise the standard of reporting so that PES administrators can calibrate their reliance on the ASPAs.

Drawing on methodological insight from data and medical science, we review the most up-to-date materials on empirical testing of ASPAs operating in the OECD and highlight the importance of reporting the false-positive error rate in establishing the operational accuracy of ASPAs (Matty, 2013).

ASPAs seek to identify those most at risk of LTU to provide targeted support to at-risk people, thus increasing efficiency and lowering costs (Desiere *et al.*, 2019). Therefore, to assess operational efficiency, we must know how (in) accurately they predict both long-term unemployment and frictional unemployment, an approach that requires multiple measures of accuracy – in particular, we need to know the error rate or how many people are incorrectly classified as high (false positive) and low risk (false negatives). Furthermore, these metrics should be provided individually because knowing the error rate in each group renders a more detailed picture of (in) accuracy, which can be termed 'operational accuracy'.

Considering error – False positives

To understand why the error rate is a crucial metric for gauging the operational effectiveness of ASPAs we need to consider the role played by humans in the profiling processes. By way of explanation, profiling algorithms function by assigning the unemployed a score for their likelihood of becoming LTU. The score is achieved by feeding (usually) hard administrative data, such as age, gender, length of unemployment, work history, qualifications etc., into the formula, generating a predictive score for LTU. However, the algorithm does not decide which scores are classified as high or low risk; this task falls to those who design and operate the technology, and this human input politicises accuracy.

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

	Size of target segment	
	Top 8%	*Top 30%*
Number of target segment that reach LTU	29	64
Number of target segment that are frictional	58	262
Total number of LTU	91	91
Total number of frictional	994	994
Proportionof all LTU captured	32%	70%
Proportion of all frictional misclassified	6%	26%

Model applied to test dataset (n=1,085)

Figure 2. Results of testing JSCI (Matty, 2013).

Choosing a cut-off point for ASPAs is remarkably important to their operation (Allhutter *et al.*, 2020)– PES have a choice to moderate the cut-off based on labour market needs or based on their own service capacity. If following the public demand for services, one would expect that ASPA operators strike the cut-off rate at a point that reflects a typical rate of LTU found in the broader population – say 8 per cent – meaning operators classify the top 8 per cent of scores predicted by the algorithm as those most likely to become LTU. Indeed, one would also expect that the cut-off point is modulated in line with macroeconomic shifts, raised as LTU rises and drops when labour markets tighten.

In practice (Rosholm *et al.*, 2004; Matty, 2013; Desiere & Struyven, 2021), when cut-off points are set to real-world levels, ASPAs have a high rate of false negatives, missing people at risk of LTU. PES tend not to worry about over-treating unemployed people with ALMPs. Indeed, they typically try to match their capacity to deliver ALMPs (for example the number of caseworkers who can deliver intensive counselling) to the cut-off, to ensure that PES work at or near capacity and that PES capacity is preserved for economic downturns. To protect the system, the cut-off threshold tends to drop when unemployment is high, and be loosened when unemployment drops. So, in practice, operators typically set the cut-off point much higher that the typical rate of LTU. We can find no evidence of cut-off points being made public, explained or reported in method statements or elsewhere, but from informal background information we have gathered, we understand ASPAs typically have cut-offs over 30 per cent, often as much as 50 per cent, reflecting the desire of PES administrators to keep ALMP capacity in use. Set at these unnaturally high levels the ASPAs correctly predict a higher number of those who become LTU (true-positives). However, as we shall see, lowering the cut-off point to increase the number of true-positives simultaneously increases the number of false-positives. An assessment of empirical studies reveals that error rates run as high as four false-positives for each true-positive (Matty, 2013). In operational terms, incredulously, this means that for each person that required extra support, PES would work with four people who did not (Matty, 2013).

Reviewing empirical testing of ASPAs

By way of demonstration, the Department of Work and Pensions (DWP) in the UK reviewed the JSCI using a definitive impact study (Matty, 2013). The method uses empirical testing of the technology in which the predictive results generated by the ASPAs are measured against real-world outcomes. In the testing process, the DWP deployed the technology on a total of 1,085 unemployed people and then followed their progress in the labour market over twelve months. The results of empirical testing are provided in Figure 2.

When set at a cut-off point of 8 per cent, as shown in Figure 3, the ASPA generates twenty-nine true positives and fifty-eight false positives. Out of ninety-one LTU individuals, twenty-nine are correctly predicted as LTU and sixty-two are incorrectly classified as low risk. Additionally, out of 994 frictionally unemployed, the tool misclassifies fifty-eight as high risk.

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

2x2 Contingency Table
Sample N=1085

	Risk Scores ↓		***LTU Group (total 91)***	***Frictional Group (total 994)***
	100 .	High Risk Group	True Positive - 29	False Positive - 58
Cut Off	 1	Low Risk Group	False Negative - 62	True Negative - 936

Figure 3. Two-by-two contingency table for predictive modelling cut off set to 8 per cent.

2x2 Contingency Table
Sample N=1085

	Risk Scores ↓		***LTU Group (total 91)***	***Frictional Group (total 994)***
	100 . . .	High Risk Group	True Positive - 64	False Positive - 262
Cut Off	 1	Low Risk Group	False Negative - 27	True Negative - 737

Figure 4. Two-by-two contingency table for predictive modelling cut off set to 30 per cent.

Therefore, in the high-risk group (top 8 per cent), there is a ratio of roughly two errors for each correct prediction making the ASPA operationally useless because, in practice, many of the people receiving extra support would not need it, and large numbers who required extra support would be denied care.

To overcome the poor performance of the ASPA, operators have the option to set a much higher cut-off point that increases the number of true positives but also increases the false-positive error rate. Figure 4 above outlines how the tool functions when the top 30 per cent of scores are classified as high risk. Set to the higher rate, the ASPA predicts more true positives – out of a total of ninety-one LTU, sixty-four (or 70 per cent) individuals are correctly classified. However, the tool simultaneously creates a much higher false-positive error rate of 26 per cent, meaning the ASPA misclassifies 262 frictionally unemployed people, as high risk. If we zone in on the high-risk group, we see that for every correct classification (true-positive), there are four incorrect classifications (false-positives). Reporting this as an accuracy rate of 70 per cent is misleading; in reality, the ASPA is operationally useless and creates significant risk for the unemployed.

Given the diversity in the ASPAs currently operating in the OECD, as a part of our evaluation, we considered the possibility that the level of error recorded by Matty, 2013 was unique to the specific algorithm. However, empirical research by the DWP in the UK (Matty, 2013) and

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

Table 2. Operational accuracy based on Matty (2013)

	Operational accuracy	
	Cut off 8%	*Cut off 30%*
Forecast accuracy rate	32%	70%
False positive rate	6%	26%
False negative error rate	68%	30%
Error ratio for high risk group	2:1	4:1
(True positive to false positive)	For every genuine LTU case, PES **misclassify two** frictionally unemployed people	For every genuine LTU case, PES **misclassify four** frictionally unemployed people

the Flemish PES (Desiere *et al., 2019)* show that model accuracy is not highly sensitive to the choice of the statistical model. In addition, empirical data available on the testing of other ASPAs such as the Belgian (Desiere & Struyven, 2021), Finnish (Riippen, 2011), Danish (Rosholm *et al.*, 2004), Irish (O'Connell *et al.*, 2009), and Swedish (Arbetsförmedlingen, 2014) suggest a high false-positive error rate is ubiquitous to all of these ASPAs.

For example, the results from empirical testing of the Danish job barometer show the total number of errors – false positives = 171,291 and false-negative = 346,270. Therefore, to correctly predict 270,953 LTU, it was necessary to misclassify 517,561 individuals or a ratio of roughly two errors to each correct prediction. Additionally, a study outlining accuracy results for the Irish PEX compares accuracy rates with the Danish job barometer and finds only a marginal improvement in the accuracy rate. Findings from the Irish study outline a false positive error rate for the PEX tool as 29 per cent, which is 3 per cent higher than the rate recorded in the UK study by Matty (2013). In the Irish case, the ASPA is used to manage the relatively fixed caseworker capacity (Roche & Griffin, 2022). Despite this, many of these empirical studies espouse the ability of ASPAs to predict the likelihood of LTU accurately.

Table 2 provides a clear and concise method of reporting the accuracy of data from the UK feasibility study conducted by Matty (2013). Reporting multiple measures of accuracy, including the rate and ratio, allows the reader to decipher the operational (in) effectiveness of ASPAs and demonstrates how misleading the current method of reporting accuracy is in terms of the operational effectiveness of ASPAs. For example, Table 2 shows that when the cut-off point is set to eight per cent the error ratio is two to one meaning that for every genuine LTU case – PES work with two people that need no intervention. Similarly, when the tool is set at a cut-off point of thirty per cent the error ratio is four to one meaning that for every genuine LTU case – PES work with four people that need no intervention.

Conclusion

This paper provides an exploratory review of reported accuracy rates for the ASPAs currently operating in the OECD as a vital first step in establishing the operational effectiveness and social impact of ASPAs and makes three contributions to our knowledge of reported accuracy rates. Firstly, reporting accuracy through a single measure is deliberately misleading and comes across as a political gesture to espouse the effectiveness of a technology that is not functioning as envisaged. Secondly, the empirical research on accuracy points to a high number of false-positive errors obfuscated by the current method of reporting accuracy as a single measure. Thirdly, providing

https://doi.org/10.1017/S1474746423000428 Published online by Cambridge University Press

multiple accuracy measures (including error rates) renders a more accurate picture of the real-world function of ASPAs and calls into question the espoused effectiveness and cost-cutting benefits of the technology and points to a negative social impact on the unemployed.

It is unclear why those who design and implement ASPAs choose to report accuracy in an opaque way. In particular, the obfuscation of a high false-positive rate suggests that ASPAs may well have undeclared benefits for key stakeholders such as PES, allowing them to manage unpredictable flows of the unemployed under chaotic labour market conditions. Alternatively, it is possible that those who espouse the effectiveness of ASPAs genuinely feel they are more accurate and efficient than other forms of profiling and accept the high false-positive rate as offering a lower deadweight cost. If this is the case, it is politically transparent why designers would choose to report accuracy in a manner that inflates ASPAs capability to predict the likelihood of LTU accurately.

What is at stake in bridging the divide between machine and policy is the effectiveness of ALMP spending, and social outcomes from such policies. It is possible that the policies policies designed to assist those most distant from the labour market are targeted, inappropriately, at the frictionally unemployed. In systems with little distinction between LTU and frictionally unemployed people's experience of ALMPs, particularly the more conditional elements, misclassification and consequential misapplication of policy measures has little import. However, in systems where the distinction is significant, being inappropriately subject to conditional ALMPs threatens individuals financial security (Lambie-Mumford, 2013) and mental well-being (Williams, 2021).

Acknowledgements. We are grateful to the participants of the European Network for Social Policy Analysis (*Espanet*) 20th anniversary conference, particularly the convenors Stefano Sacchi, Minna van Gerven and Magnus Paulsen Hansen and the reviewers for their editorial suggestions.

Financial support. Supported by funding from the Horizon 2020 Framework Programme, Grant/Award Number: 870702; EU Framework Programme for Research and Innovation.

References

Allhutter, D., Cech, F., Fischer, F., Grill, G. and Mager, A. (2020) 'Algorithmic profiling of job seekers in Austria: How austerity politics are made effective,' *Big Data*, **3**, 5. https://doi.org/10.3389/data

Arbetsförmedlingen (2014) 'Arbetsförmedlingens Återrapportering: Insatser för att förhindra långvarig arbetslöshet', Arbetsförmedlingen Reports 2014: Efforts to prevent long-term unemployment.

Bader, V. and Kaiser, S. (2019) 'Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence,' *Organization* **26**, 5, 655–672.

Behncke, S., Frölich, M., and Lechner, M. (2007) *Public employment services and employers: how important are networks with firms?* Bonn: Institute for the Study of Labor. https://www.econstor.eu/bitstream/10419/34342/1/54557224X.pdf

Bejaković, P. and Mrnjavac, Ž. (2018) 'The danger of long-term unemployment and measures for its reduction: The case of Croatia,' *Economic Research-Ekonomska istraživanja*, **31**, 1, 1837–1850.

Boland, T., and Griffin, R. (Eds.). (2015) *The Sociology of Unemployment*. Manchester University Press.

Bonoli, G. (2010) 'The political economy of active labor-market policy,' *Politics and Society*, **38**, 4, 435–457.

Botrić, V. (2017) *LTU Recommendation implementation in Croatia*. Zagreb: Presentation.

Brandt, M. and Hank, K. (2014) 'Scars that will not disappear: Long-term associations between early and later life unemployment under different welfare regimes,' *Journal of Social Policy*, **43**, 4, 727.

Card, D., Kluve, J. and Weber, A. (2018) 'What works? A meta analysis of recent active labor market program evaluations,' Journal of the European Economic Association, **16**, 3, 894–931.

Danaher, J. (2016) 'The threat of algocracy: Reality, resistance and accommodation,' Philosophy and Technology, **29**, 3, 245–268.

Demazière, D. (2014) 'Does unemployment still have a meaning? Findings from a comparison of three conurbations,' *Sociologie du travail*, **56**, e21–e42.

Desiere, S., Langenbucher, K. and Struyven, L. (2019) 'Statistical profiling in public employment services: An international comparison,' pp. 1–29. OECD. https://doi.org/10.1787/b5e5f16e-en [accessed 11.02.2020].

Desiere, S. and Struyven, L. (2021) 'Using artificial intelligence to classify job seekers: The accuracy-equity trade-off,' *Journal of Social Policy*, **50**, 2, 367–385.

Ferrante di Ruffano, L., Hyde, C. J., McCaffery, K. J., Bossuyt, P. M. and Deeks, J. J. (2012) 'Assessing the value of diagnostic tests: A framework for designing and evaluating trials,' *British Medical Journal (Clinical Research ed.)*, *344*, e686. https://doi.org/10.1136/bmj.e686

Fleischer, K. (2016) *Statistically Assisted Profiling - Client Support by Appropriate Tools*. Zagreb: Presentation.

Georges, N. (2008) Le profilage statistique est-il l'avenir des politiques de l'emploi? *L'emploi, nouveaux enjeux*, 117–124.

Goffman, E. (1959) *The Presentation of Self in Everyday Life*. New York: Anchor Books.

Griffin, R., Boland, T., Tuite, A. and Hennessy, A. (2020) 'Electric dreams of welfare in the 4th industrial revolution: An actor-network investigation and genealogy of an algorithm,' In *Digitisation and Precarisation* (pp. 181–203). Wiesbaden: Springer VS.

Hasluck, C. (2008) 'The use of statistical profiling for targeting employment services: The international experience,' in G. Di Domenico and S. Spattini (eds.), *New European approaches to long-term unemployment: What role for public employment services and what market for private stakeholders?* Kluwer Law International BV.

Kappen, T. H. and Peelen, L. M. (2016) 'Prediction models: The right tool for the right problem,' *Current Opinion in Anaesthesiology*, **29**, 6, 717–726.

Kappen, T. H., van Klei, W. A., van Wolfswinkel, L., Kalkman, C. J., Vergouwe, Y. and Moons, K. G. (2018) 'Evaluating the impact of prediction models: Lessons learned, challenges, and recommendations,' *Diagnostic and Prognostic Research*, **2**, 1, 1–11.

Kaufman, J. (2020) 'Intensity, moderation, and the pressures of expectation: Calculation and coercion in the street-level practice of welfare conditionality,' *Social Policy and Administration*, **54**, 2, 205–218.

Lambie-Mumford, H. (2013) "Every town should have one': Emergency food banking in the UK,' *Journal of Social Policy*, **42**, 1, 73–89.

Larsen, A. and Jonsson, A. (2011) 'Employability profiling systems – The Danish experience,' in *Presentation, Public Employment Services Conference*.

Lipp, R. (2005) Job seeker profiling: The Australian experience. In *EU-Profiling Seminar*.

Lipsky, M. (2010) Street-Level Bureaucracy: Dilemmas of the Individual in Public Service. Russell Sage Foundation.

Loopstra, R. and Tarasuk, V. (2015) 'Food bank usage is a poor indicator of food insecurity: Insights from Canada,' *Social Policy and Society*, **14**, 3, 443–455. https://doi.org/10.1017/S1474746415000184

Lopez, P. (2019) 'Reinforcing intersectional inequality via the AMS algorithm in Austria,' in *Critical Issues in Science, Technology and Society Studies. Conference Proceedings of the STS Conference (Graz: Verlag der Technischen Universität)* (pp. 1–19).

Loxha, A. and Morgandi, M. (2014) 'Profiling the unemployed: S review of OECD experiences and implications for emerging economies,' *Social Protection and labour discussion paper*. Washington, DC: World Bank Group.

Lustig, C. and Nardi, B. (2015) 'Algorithmic authority: The case of Bitcoin,' in *2015 48th Hawaii International Conference on System Sciences* (pp. 743–752). IEEE.

Martin, J. P. and Grubb, D. (2001) 'What works and for whom: A review of OECD countries' experiences with active labour market policies,' *Swedish Economic Policy Review*, **8**, 14, 9–56.

Matty, S. (2013) 'Predicting likelihood of long-term unemployment: The development of a UK jobseekers' classification instrument,' *Corporate Document Services*.

McDonald, C., Marston, G. and Buckley, A. (2003) 'Risk technology in Australia: The role of the job seeker classification instrument in employment services,' *Critical Social Policy*,**23**, 4, 498–525.

McGuinness, S., Kelly, E. and Walsh, J. R. (2014) 'Predicting the probability of long-term unemployment in Ireland using administrative data,' *Economic and Social Research Institute (ESRI) Research Series*.

Ministry of Social Development. (2018) *Implementation Plan: Client Service Matching Effectiveness Model*. Wellington City: Ministry of Social Development.

Niklas, J., Sztandar-Sztanderska, K., Szymielewicz, K., Baczko-Dombi, A. and Walkowiak, A. (2015) 'Profiling the unemployed in Poland: Social and political implications of algorithmic decision making,' *Fundacja Panoptykon, Warsaw Google Scholar*.

O'Connell, P., McGuinness, S., Kelly, E. and Walsh, J. (2009) *National Profiling of the Unemployed in Ireland*. Dublin: ESRI. https://www.esri.ie/system/files/media/file-uploads/2015-07/RS010.pdf [accessed 14.03.2020].

OECD (2019a) *Strengthening Active Labour Market Policies in Italy, Connecting People with Jobs*. Paris: OECD Publishing. https://doi.org/10.1787/160a3c28-en

OECD (2019b) *Evaluating Latvia's Active Labour Market Policies, Connecting People with Jobs*. Paris: OECD Publishing. https://doi.org/10.1787/6037200a-en

Owalgroup (2019) Artificial Intelligence in Employment Services - A mapping. https://tem.fi/documents/1410877/15020328/Artificialintellingence in employment services-Amapping/24844ede-0570-c8da-4ed3-c91ec25b8e76/Artificialintellingence inemploymentservices-Amapping.pdf

Pignatti, C. and Van Belle, E. (2021) 'Better together: Active and passive labor market policies in developed and developing economies,' IZA Journal of Development and Migration, **12**, 1.

Ponomareva, N., and Sheen, J. (2013) 'Australian labor market dynamics across the ages,' *Economic Modelling*, **35**, 453–463.

Reilly, B. and Evans, A. (2006) 'Translating clinical research into clinical practice: Impact of using prediction rules to make decisions,' *Annals of Internal Medicine*, 144, 201–209.

Riipinen, T. (2011) Risk profiling of long-term unemployment in Finland. In *Power Point Presentation at the European Commission's "PES to PES Dialogue Dissemination Conference,"* Brussels, September (pp. 8–9).

Roche, Z. and Griffin, R. (2022) Activation through Marketisation as a Process of Ignorancing. Social Policy and Administration.

Rosholm, M., Svarer, M. and Hammer, B. (2004) A Danish Profiling System (pp. 1–24). Bonn: Institute for the Study of Labor. http://ssrn.com/abstract = 628062

Sage, D. (2015) 'Do active labour market policies promote the subjective well-being, health and social capital of the unemployed? Evidence from the UK, *Social Indicators Research*, **124**, 2, 319–337.

Senghaas, M., Freier, C. and Kupka, P. (2019) 'Practices of activation in frontline interactions: Coercion, persuasion, and the role of trust in activation policies in Germany,' *Social Policy and Administration*, **53**, 5, 613–626.

Strandh, M. and Nordlund, M. (2008) 'Active labour market policy and unemployment scarring: A ten-year Swedish panel study,' *Journal of Social Policy*, **37**, 3, 357–382.

Sundberg, T. (2017) 'Systematic reviews in social policy evaluation,' in *Handbook of Social Policy Evaluation*. Edward Elgar Publishing.

Swets, J. A. (1988) 'Measuring the accuracy of diagnostic systems,' *Science*, **240**, 4857, 1285–1293.

Tashman, L. J. (2000) 'Out-of-sample tests of forecasting accuracy: An analysis and review,' *International Journal of Forecasting*, **16**, 4, 437–450.

Van Landeghem, B., Desiere, S. and Struyven, L. (2021) 'Statistical profiling of unemployed jobseekers: The increasing availability of big data allows for the profiling of unemployed jobseekers via statistical models,' *IZA World of Labor*.

Wijnhoven, M. A., and Havinga, H. (2014) 'The work profiler: A digital instrument for selection and diagnosis of the unemployed. *Local Economy*, **29**(6–7), 740–749. https://doi.org/10.1177/0269094214545045

Williams, E. (2021) 'Unemployment, sanctions, and mental health: The relationship between benefit sanctions and antidepressant prescribing,' *Journal of Social Policy*, **50**(1), 1–20.

Wollmann, H. (2003) 'Evaluation in public-sector reform. Towards a "third wave" of evaluation,' in *Evaluation in Public-Sector Reform* (pp. 1–11). Cheltenham, UK: Edward Elgar.

Zejnilović, L., Lavado, S., Martínez de Rituerto de Troya, Í., Sim, S. and Bell, A. (2020) 'Algorithmic long-term unemployment risk assessment in use: Counselors' perceptions and use practices,' *Global Perspectives*, **1**, 1.

Cite this article: Gallagher P and Griffin R. (in) Accuracy in Algorithmic Profiling of the Unemployed – An Exploratory Review of Reporting Standards. *Social Policy and Society*. https://doi.org/10.1017/S1474746423000428