OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 0305-9049
doi: 10.1111/obes.12550

# Using Machine Learning to Create an Early Warning System for Welfare Recipients*

DARIO SANSONE†,‡ and ANNA ZHU‡,§

†*Department of Economics, University of Exeter Business School, University of Exeter, Rennes Drive, Exeter EX4 4PU, UK (e-mail: d.sansone@exeter.ac.uk)*
‡*IZA, Bonn, Germany (e-mail: anna.zhu@rmit.edu.au)*
§*RMIT University, Melbourne, Victoria, Australia*

## Abstract

Using high-quality nationwide social security data combined with machine learning tools, we develop predictive models of income support receipt intensities for any payment enrolee in the Australian social security system between 2014 and 2018. We show that machine learning algorithms can significantly improve predictive accuracy compared to simpler heuristic models or early warning systems currently in use. Specifically, the former predicts the proportion of time individuals are on income support in the subsequent 4 years with greater accuracy, by a magnitude of at least 22% (14 percentage points increase in the R-squared), compared to the latter. This gain can be achieved at no extra cost to practitioners since the algorithms use administrative data currently available to caseworkers. Consequently, our machine learning algorithms can improve the detection of long-term income support recipients, which can potentially enable governments and institutions to offer timely support to these at-risk individuals.

## I. Introduction

Long-term income support (welfare) receipt is an issue many governments around the world aim to prevent (HM Government, 2010; Welfare Working Group, 2011; Reddel, 2018; Scoppetta and Buckenleib, 2018; Hanna, 2019). In basic security and/or targeted welfare systems, income support payments are designed to provide a minimum standard of living to households who are unable to meet essential consumptions needs with income from private sources (Korpi and Palme, 1998). Individuals who regularly receive income support – over an extended period of time – are most likely to suffer

JEL Classification numbers: C53, H53, I38, J68.

*We thank the Editor, two anonymous referees, Bruce Bradbury, Simon Feeny, David McKenzie, Tim Reddel, and Tim Robinson, as well as the participants of seminars delivered at the Social Policy Research Centre, University of New South Wales, the Australian National University, University of Exeter, and IMT Lucca for their helpful comments. Yin King Fok provided excellent research assistance. Zhu acknowledges the support of the Australian Research Council (ARC) Linkage Project (LP170100472). This paper uses unit record data from the Centrelink administrative records from the Department of Social Services (DSS). The findings and views reported in this paper are those of the authors and should not be attributed to the ARC or DSS. All errors are our own.

long-term economic disadvantage and social exclusion (Whiteford, 2010). Entrenched reliance on income support also imposes significant demands on government budgets, reduces economy-wide market output, and leads to the intergenerational transmission of welfare cultures (Dahl, Kostøl, and Mogstad, 2014; Dahl and Gielen, 2021; Cobb-Clark *et al.*, 2022). Individuals remaining unemployed for extended periods is an ongoing issue given the significant changes to the economy from automation, technological change, and the continuing impacts of the COVID-19 pandemic.[1]

To prevent such entrenched reliance, policymakers try to intervene early in the welfare careers of high-risk registrants with labour market activation programmes, further supports, and casework management. Interventions that provide tailored support and recognise an individual's specific needs and constraints have been shown to improve employment prospects in both the short-term and long-term (Card, Kluve, and Weber, 2018; Vooren *et al.*, 2019). A first step in early intervention programmes, however, is to identify a target group. In Australia, for example, the government is specifically focusing on targeting individuals who are the most disadvantaged and thus most likely to stay on welfare for a protracted period of time (Department of Social Services, 2018; Department of Employment, Skills, 2020). Currently, identifying these at-risk individuals involves simple profiling tools, valuations of lifetime welfare costs, and/or caseworkers' evaluations.

Our aim is to evaluate whether Machine Learning (ML) tools combined with high-quality nationwide administrative data can improve upon existing screening practice. In particular, we develop predictive ML models of income support receipt intensities for any payment enrolee in the Australian social security system. A key attraction of an Australian case-study is that it has one of the most targeted welfare systems in the world, which means that when we predict the risk of long-term income support receipt, we closely predict the incidence of ongoing poverty and social exclusion (Whiteford, 2010).

*Ex-ante*, it is unclear whether supervised ML algorithms applied to the available administrative data will outperform simpler models or currently used early warning systems in the Australian social welfare system. Certainly, the literature documents many cases where ML has failed to outperform simpler models (Beattie, Laliberte, and Oreopoulos, 2018; McKenzie and Sansone, 2019; Salganik *et al.*, 2020; Orlov *et al.*, 2021). Limited predictive performance can arise because researchers have insufficient number of observations or miss out the relevant inputs to properly calibrate a ML algorithm. If, for instance, a certain outcome variable is mainly driven by unobservable characteristics, then adding nonlinearities or high-order interaction terms among observable characteristics would fail to substantially enhance the performance of a ML algorithm (McKenzie and Sansone, 2019). Similarly, if the true underlying relationship among some variables is linear, then using more flexible functional forms will not lead to more accurate predictions.

One of our key contributions to the literature is to *ex-post* show that ML models, when applied to novel administrative data, do significantly improve predictive accuracy compared to simpler predictive models or existing early warning systems in use.

[1] It is worth noting that welfare receipt is not necessarily a problem in and of itself since it can serve to protect the health and well-being of recipients, particularly those who face unexpected shocks or major structural impediments to work (Aizer and Currie, 2004; Mitrut and Tudor, 2018). In contrast, entrenched reliance on income support is considered an unintended cost of work disincentives embedded in program designs, such as that of highly targeted welfare systems (Moffitt, 1985, 1992; Hoynes, 1997; Feldstein, 2005; Hoynes and Schanzenbach, 2012).

Specifically, our simulations show how ML algorithms can increase the R-squared of models predicting the proportion of time individuals will be on income support by at least 22% (or 14 percentage points) compared to alternative models. As a result, our ML algorithms can improve the detection of long-term income support recipients accruing a welfare cost nearly 1 billion AUD higher than individuals identified in the current system. This is an important contribution for two reasons. First, until now, we have had limited understanding about the capacity of ML models to make better predictions in this domain. Second, our research is policy relevant and timely since the Australian government is currently trialling innovative profiling tools in the targeting of early intervention programmes with the specific aim of preventing welfare dependency (Reddel, 2018). From an equity angle, individuals who are at-risk of receiving income support for prolonged periods of time are likely to belong to some of the most vulnerable groups in society, especially given Australia's highly targeted income support system. Thus, identifying these at-risk individuals can also enable policymakers to better support them with additional resources and services. If proven successful, this early warning system could thus act as a blueprint for other countries, many of whom are already exploring the use of Artificial Intelligence in social protection programmes (Lokshin and Umapathi, 2022), have similar welfare system, and comparable data infrastructure.

Another contribution of this paper is the use of novel, high-quality administrative data. We use a dataset called 'Data On Multiple INdividual Occurrences' (DOMINO), which includes the full population (over 32 million persons) who had contact with the Social Security System. These data have several key benefits. First, they are high frequency with daily information of income support receipt status over the entire analysis period. These repeat observations enable an analysis of the dynamics of welfare exit and entry. Second, they include both welfare recipients and individuals receiving other, more universal government transfers such as family payments. Thus, we can circumvent issues of non-up-take of income support benefits by eligible individuals. Third, the data set is large and is rich in covariates (approximately 1,800 possible predictors), which makes it ideal for calibrating ML algorithms. Fourth, the data are high quality because the government relies on these exact data to determine an individual's eligibility to income support payments. Since an individual's payment amount is a direct function of their income, wealth, savings, household structure, and several other socio-economic factors, the information in these data is reconciled with Australian Tax Office records to ensure accuracy. Last but not least, in contrast with previous ML applications relying on survey data and/or data that are not directly available to practitioners or relevant to their day-to-day tasks (Van Landeghem, Desiere, and Struyven, 2021), caseworkers and front-line agency staff already have access to these administrative data. This allows us to validate our algorithms with data that are representative of the population of interest and makes it easier to integrate the ML algorithms developed here in future decision-making processes. There is also evidence showing that administrative data tend to outperform behavioural data and lessen privacy-invasion concerns (Bjerre-Nielsen *et al.*, 2021).

We argue that our estimated ML models applied to these novel and rich data can be a large part of the solution to the resource allocation problem faced by welfare agencies. However, we acknowledge the limitations of ML, particularly when applied to a population of already-vulnerable individuals such as long-term welfare recipients.

Automated systems and ML models have been charged with de-individualising people, discriminating against minority groups, and perpetuating systemic inequalities intrinsic in the data (Madhusoodanan, 2021; Staines *et al.*, 2021). Furthermore, they are devoid of theory and may have limited capacity to uncover causal impacts. This means we cannot address important policy questions such as 'which' interventions work or 'who' are most responsive to them.

Predictions in this paper are indeed different from treatment effect estimates that a policymaker may use to optimally allocate resources (Athey, 2017). Yet, while our paper is focused on *prediction*, it can be seen as a first step to identify and characterise at-risk individuals. For example, in a follow-up phase, caseworkers could focus on a particular sub-group and use their expertise (potentially in combination with contextual bandit algorithms or findings from previous RCTs) to assign these individuals to specific programmes with proven high treatment *causal* effects. Moreover, ML can further supplement causal analyses by identifying subgroups based on complex interactions between background characteristics. This can improve upon existing analyses that focus on one dimension to form these subgroups (e.g. age, sex or race), and thus provide the basis for examining more intricate and multidimensional heterogeneous treatment effects in causal analyses. We therefore believe, as also advocated in other ML applications (Hofman *et al.*, 2021), that this paper provides an example of how ML can be used to obtain accurate predictions and complement – rather than substitute – standard econometric techniques aimed at estimating causal effects. In addition, from a social welfare perspective, policymakers may prioritise allocating resources to the most vulnerable individuals even if it does not necessarily aim to maximise the marginal gains from a treatment programme.

Within this context, and after having identified at-risk individuals, we apply unsupervised ML to cluster such individuals into different groups based on their observable characteristics. Clustering at-risk individuals emphasizes that these individuals are not a homogeneous group. Indeed, the ML algorithm may classify some individuals as at-risk because of their disabilities, or because they have caring responsibilities, while others may be predicted to be at risk because of their migration status or age. The latter group would likely require different intervention programmes from the former. This application is therefore an example of how ML can be used to identify and group individuals at-risk and complement subsequent causal analyses in designing treatments appropriate for each sub-population.

Furthermore, we argue the predictive models can improve decision-making by systemizing the process of identification. In fact, it can potentially detect and thus prevent conscious and/or unconscious biases by caseworkers (Kretsedemas, 2005; McBeath *et al.*, 2014; Pentaraki, 2019; Kleinberg *et al.*, 2020). Importantly, in this context the possibility that ML may predict certain subgroups such as indigenous people to be more likely to stay on income support for an extended period of time could actually benefit these individuals if caseworkers can then target and properly support them (Desiere and Struyven, 2021).

On a practical level, our ML models can help to promptly identify at-risk individuals using only short-term variables, potentially even as soon as they register with the system. This is especially useful when there has been no prior familiarity with the registrant's situation. In this way, our predictive models can reduce workload pressures on caseworkers (because they act as automatic screening devices) and can avoid an arbitrary selection of predictors or subgroups. Current practices to identify high-risk registrants over-burden caseworkers: for example, in Australia, a caseworker at any one time has an average caseload of more than 100 clients (Davidson, 2019). This labour-intensive screening process may subsequently divert resources away from individualised intervention efforts. Furthermore, frontline staff may make 'cream-skimming' targeting decisions, such as skewing the offering of employment activation programmes towards those for whom outcomes are easier to achieve. This is especially prevalent for countries that contract out a high percentage – 100% in Australia's case – of employment services to private agencies and where compensation is tied to outcomes (O'Sullivan, McGann, and Considine, 2019). Similarly, in contrast with heuristic evaluations and predictions relying solely on caseworker's expertise and training, these ML predictions can be provided to caseworkers without requiring them to undergo any extra training on ML. Last but not least, these improvements can be made at low-cost. In effect, when our predictive models are paired with caseworker expertise, they can enable a superior assignment policy.

Our focus is on the outcome of any income support receipt since this is also the focus of the Australian government. Given its emphasis in the past economic literature (Card, Chetty, and Weber, 2007; Card *et al.*, 2015; Schmieder and von Wachter, 2016), we then estimate – as an extension – ML algorithms for the outcome of long-term receipt of unemployment-specific benefits. Unemployment benefit recipients are a subset of all income support recipients. A distinguishing feature of benefit eligibility for jobseekers is that they have additional job search and activity requirements to meet, compared to a recipient of other types of income support payment such as disability benefits or parenting allowances.

The topic of our research speaks to two previous bodies of work. The first is the nascent literature on estimating ML models to inform resource allocation decisions. This literature has shown that ML algorithms can assist personnel with day-to-day decisions because they better predicts at-risk individuals compared to standard regression tools (Kleinberg *et al.*, 2015). For example, judges can improve bail-granting decisions (Kleinberg *et al.*, 2017), health inspectors can be more efficiently allocated to unhygienic restaurants (Kang *et al.*, 2013; Glaeser *et al.*, 2018), programme administrators can better target interventions at high-risk youth (Chandler, Levitt, and List, 2011), employers can make better hiring decisions (Hoffman, Kahn, and Li, 2018), schools can promptly identify students at risk of dropping out (Sansone, 2019), government officials can prevent child abuse and maltreatment (Vaithianathan *et al.*, 2013; Cuccaro-Alamin *et al.*, 2017), resettlement agency staff can improve short-run employment outcomes of refugees (Ahani *et al.*, 2021), and surgeons can more ethically and effectively screen patients for hip-replacement surgery (Mullainathan and Spiess, 2017). In addition, researchers have combined ML with satellite data to predict poverty levels in countries with limited survey and administrative data (Jean *et al.*, 2016; Yeh *et al.*, 2020; Huang, Hsiang, and Gonzalez-Navarro, 2021). Researchers have also started to analyse how (and if) these algorithms can add value when they are used to complement the judgement and skills of human experts (Stevenson and Doleac, 2019; Ibrahim, Kim, and Tong, 2021). For example, once ML models identify high-risk individuals, caseworkers can then choose which types (or intensities) of programmes ought to be targeted at which individuals – one way in which to address the contextual bandit problem (Athey, 2019). We innovate in the ML literature by being one of the first scholars to develop predictive ML models of long-term income support receipt.

A second literature is the one that examines the factors explaining long-term income support receipt. Several relevant issues emerge when analysing this topic and they often stem from data limitations. First, welfare churn (i.e. repeated exit and re-entry into the welfare system) and welfare scarring (i.e. the current probability of income support receipt increasing as a result of income support receipt in the past) are common phenomena: unlike our high-frequency daily data, point-in-time measures or yearly snapshots of welfare receipt can fail to capture these dynamics (Barrett, 2000; Tseng and Wilkins, 2003; Bäckman and Bergmark, 2011; Bhuller, Brinch, and Königs, 2017). Second, non-take-up of income support payments among eligible individuals is a significant issue and using data that excludes welfare non-recipients – in contrast with our data including universal government transfers recipients – fails to shed light on their income and employment outcomes (Currie and Grogger, 2001; Bitler, Currie, and Scholz, 2003). And last, the influence of personal characteristics (such as age, sex, ethnicity, household structure) has been shown to interact with the impact of structural factors (such as geography, policy reform, organizational incentives, macroeconomic conditions) in explaining entrenched welfare receipt (Leana, Mittal, and Stiehl, 2012; Bradbury and Zhu, 2018). Therefore, accurately and flexibly modelling relationships between individual characteristics and long-term income support receipt requires larger data combined – through ML models – with a richer set of covariates. We innovate in this literature by using a novel, high-quality administrative dataset, which allows us to address the empirical issues mentioned above.

## II. Institutional background

The primary purpose of Australia's social security system is to provide people with a 'minimum adequate standard of living' (Australian Treasury, 2010). The main class of benefits provided are called 'income support payments' and they are targeted at individuals with no or low levels of income and/or assets. Generally, for these recipients, income support payments serve as their primary source of income. These welfare payments are provided on a regular basis and assist with basic living costs. The maximum annual income support amount in 2018 ranged from 12,400 AUD (for unemployment benefits) to 20,700 AUD (for disability benefits), whereas the median annual personal income was 48,400 AUD (ABS, 2019b). Thus, those who are receiving an income support payment are a highly disadvantaged group.

There are six main categories of income support payments: (1) student payments; (2) unemployment payments; (3) parenting payments; (4) disability payment; (5) carer payment; and (6) age pension. These main income support payments are strictly means-tested. This means that a formal process is used to determine an individual's eligibility for payments. Entitlement is based on current (not previous) levels of income and assets. Cash transfer amounts reduce when earnings and assets increase. This targeted model applies to all the income support payments; however, the income thresholds and taper rates (i.e. how much welfare transfers decrease when earning and assets increase) vary depending on the type of income support payment. For example, Unemployment benefits have one of the highest taper rates at 50%–60%, which means that benefits cut out as soon as recipients earn roughly 24,400 AUD. The taper rates for other benefits are lower at 40%–50%. Unemployment benefits in Australia are flat-rate payments provided to any individual who is currently unemployed, conditional on them satisfying activity test requirements.[2] It is important to note that unemployment benefits are distinct from unemployment insurance because eligibility for the former does not depend on previous earnings.

This strictness of the income and asset tests gives the Australian welfare system the title as one of the most targeted systems in the OECD. For example, in 2005, the average share of transfers received by the poorest population quintile as a ratio of the share received by the richest quintile was 2.1 across the OECD but as high as 12.1 for Australia. In other words, 'the poorest 20 per cent of the [Australian] population [received] more than 12 times as much in social security benefits as the richest quintile' (Whiteford, 2010).

For the most part, the welfare system aims to support people in immediate need. It does this by providing highly targeted income support payments. Nearly a fourth of the Australian population receive these payments. However, the Australian government also provides financial support through the social security system that is not considered an income support payment. Other payments available through the social security system include: payments that are designed to assist families with the cost of raising children, such as family tax benefits, maternity leave benefits, supplementary payments for the main income support benefits; and rent assistance. As some of the non-income support payments are relatively untargeted, the social security system includes cross-sections of the population that are not necessarily financially disadvantaged. As an example of the composition of recipients in the social security system, in any one fortnight in 2018, about 5 million Australians received an income support payment while approximately 855,000 families received family tax benefits (Whiteford, 2018). In total, social security payments equalled 112.4 billion AUD in the 2017–18 fiscal year (AIHW, 2019b).

While the Australian social security system provides for a minimum standard of living with payments targeted to people who do not have the means to support themselves, this goal is balanced with the aim to encourage self-reliance and reciprocity by including activity tests, monitoring, and sanctions (Klapdor, 2013). For example, to remain qualified for unemployment payments, recipients are required to actively look, plan, and prepare for work in the future. Failure to comply with these mutual obligation requirements can result in loss of payments. Sanctions represent a punitive aspect of mutual obligations and many have argued that the penalties for noncompliance are disproportionate and counter-productive (Australian Senate Committee on Community Affairs, 2021; UnitingCare Australia, 2021). Yet, there are also aspects of the mutual obligations programme that can be supportive. For example, specialised services exist for jobseekers with very young children, disability, or living in remote area (Australian Senate Committee on Community Affairs, 2021), and jobseekers have access to targeted assistance in work-ready skills such as additional training opportunities, interview skills, help with resumes, and referrals to appropriate work opportunities.

[2] New migrants are ineligible for unemployment benefits within four years of arriving in the country.

Currently, the government contracts out all employment assistance programmes for unemployed people to for-profit and not-for-profit providers under the ‘Job Services Australia’ programme. Providers are paid according to a combination of service inputs, investment in work experience and training, and importantly, funding is linked to employment outcomes. Researchers who have evaluated the Job Services Australia programme have found evidence of ‘cream-skimming’ by caseworkers, such as targeting employment activation programmes at those for whom outcomes are easier to achieve – since payments for providers are tied to outcomes (O’Sullivan *et al.*, 2019). As mentioned in the introduction, ML models, by providing risk scores for each individual, can provide detailed guidance to the government about which individuals are most at-risk of long-term unemployment. This information can then inform payment structures for contracted-out employment services. In this case, ML can serve as a critical governance and auditing tool for government-employment service provider relations.

The government has also additional education and training programmes and mentoring services specifically aimed at reducing long-term income support receipt. For example, in Australia in 2016, as part of a 92-million-dollar initiative called ‘Try, Test and Learn,’ the government targeted early intervention programmes at three high-risk groups of individuals: young carers aged 24 or under, young parents aged 18 or under, and students receiving study-related income support payments. These groups were selected based on an actuarial valuation of their estimated lifetime welfare costs (Department of Social Services, 2018, Price Waterhouse Coopers, 2016; 2019). Other considerations included the adequacy of available support services for these groups and their expected responsiveness to new models of policy interventions. The initiative funded one-on-one tailored support, including short-term education, workshops, mentoring, work placements and self-employment/business start-up initiatives. In this context, ML models can help to further refine the process of identification of high-risk individuals, especially those within the three identified broad groups who may have different risk profiles. Importantly, no other programme targeted the same groups in 2014, that is, in the year used to compute our baseline inputs.

## III. Data

### Population

Australian federal social security records from 2000 to 2019 are available in the DOMINO dataset. All social security payments are administered by the national welfare agency called Centrelink. There are over 32 million persons in these data who had any contact with the Centrelink system between 2000 and 2019. All registrants are 15 years old or above as this is the minimum age of eligibility. The financial circumstances of these registrants vary greatly: some have high levels of financial needs, such as those who are in receipt of targeted income support payments; others have higher incomes and register with the social security system because they receive one of the non-income support payments described in the previous section, such as one-off government bonuses or cost-of-children payments.

Each individual is tracked over time on a highly frequent basis. For our main variable of welfare receipt status, we know the precise start and end date associated with payment receipt. A key advantage of this data structure for our study is that we can construct a precise picture of the duration and dynamics of welfare receipt over a long period of time.

The reason we observe such high-(daily)-frequency data is because income support (or welfare) payments are highly targeted. This means that recipients' eligibility for payments are assessed regularly, and an array of information is collected about the recipients. To begin, all recipients must fill in a 14- to 34-page form that elicits information about the recipients' (and if applicable, partners') basic information (name, address, contact details, gender, date of birth, ethnicity, language, citizenship, arrival information), marital status and relationship event history, demographic information about their dependent children, accommodation details, employment and study details. Thereafter, all income support recipients are expected to update Centrelink on their income, living arrangements and financial circumstances on a regular (bi-weekly) basis. Recipients are also required to report changes (such as to relationship status, employment outcomes, changes in wealth or living conditions) within 14 days of the change. The start and end dates of payment receipt are then recorded in our data.

## Sample

Our first step is to identify the universe of individuals who received any type of payment from Centrelink in 2014, and who were aged 15–66 on 1 January 2014. Although the full data cover the period 2000–19, we have chosen 2014 as our base year to calculate our input variables. We have then calculated an individual's welfare receipt intensity from 2015 to 2018, inclusive. Note that we have ignored the data from 2000 to 2013. This allows us to estimate the ML algorithms in a period after the Global Financial Crisis effects have washed out. We have also chosen to exclude the year of 2019 because the end of our observation window is 14 October 2019. Partial inclusion of 2019 data points could bias our results if there are strong seasonal patterns associated with welfare receipt. We have chosen to focus on the working-age population since this is the target group for programmes aimed at preventing long-term welfare dependency. Our next step is, for computational reasons driven by the power of the server where the data are stored, to draw a random 1% sample (50,615 individuals) from this population.[3]

The resulting sample is not representative of the whole Australian population (Table 1). It disproportionately captures individuals with low- to middle-incomes. The highly targeted nature of Australian welfare programmes means low-income individuals are over-represented. Furthermore, the non-income support payments, such as the cost-of-children payments, are provided to primary carers of children. This means that there is a higher percentage of female individuals in our sample since they continue to be over-represented among primary carers of children (as well as among individuals eligible for other carer-related payments). Nevertheless, it is worth emphasizing that this sample is appropriate for our goal of predicting welfare dependency given its focus on vulnerable individuals.

[3] Note that the Australian population aged above 15 in 2014 was approximately 19 million (ABS, 2014). Around 7 million individuals were registered in the Centrelink system in 2014 (our base year). Of the 2014 Centrelink registrants, roughly 5 million were aged 15–66 years. A 1% sample gives us our final sample of 50,615 individuals.

TABLE 1

*Summary statistics comparison between Census and DOMINO data*

| | *Census* | *DOMINO* |
|---|---|---|
| Demographics | | |
| Age (mean) | 39.825 | 40.954 |
| Female | 0.505 | 0.620 |
| Ever a parent (with at least one child) | 0.291 | 0.468 |
| Indigenous | 0.027 | 0.057 |
| Australian-born | 0.680 | 0.721 |
| Immigrant | 0.320 | 0.279 |
| Education | | |
| Year 12 or below | 0.415 | 0.549 |
| Certificate 1–2 or below | 0.416 | 0.785 |
| Certificate 3–4 or below | 0.619 | 0.839 |
| Diploma or below | 0.725 | 0.904 |
| Bachelor’s degree or below | 0.915 | 0.986 |
| Marital status | | |
| Single | 0.430 | 0.542 |
| Married | 0.431 | 0.361 |
| De Facto | 0.133 | 0.097 |
| Separated | 0.025 | 0.182 |
| Divorced | 0.058 | 0.038 |
| Widowed | 0.011 | 0.018 |
| State of residence | | |
| New South Wales | 0.318 | 0.318 |
| Victoria | 0.255 | 0.248 |
| Queensland | 0.200 | 0.211 |
| Western Australia | 0.070 | 0.087 |
| South Australia | 0.107 | 0.080 |
| Australian Capital Territory | 0.021 | 0.012 |
| Northern Territory | 0.011 | 0.011 |
| Tasmania | 0.018 | 0.027 |
| Weekly personal income (annual income) | | |
| Nil (or negative) income | 0.106 | 0.134 |
| \$1–\$149 (\$1–\$7,799) | 0.046 | 0.467 |
| \$150–\$299 (\$7,800–\$15,599) | 0.067 | 0.134 |
| \$300–\$399 (\$15,600–\$20,799) | 0.059 | 0.052 |
| \$400–\$499 (\$20,800–\$25,999) | 0.060 | 0.021 |
| \$500–\$649 (\$26,000–\$33,799) | 0.069 | 0.012 |
| \$650–\$799 (\$33,800–\$41,599) | 0.077 | 0.010 |
| \$800–\$999 (\$41,600–\$51,999) | 0.089 | 0.005 |
| \$1,000–\$1,249 (\$52,000–\$64,999) | 0.093 | 0.003 |
| \$1,250–\$1,499 (\$65,000–\$77,999) | 0.065 | 0.001 |
| \$1500 or more (\$78,000 or more) | 0.187 | 0.000 |
| Not stated | 0.082 | 0.295 |

*Notes*: Samples for both Census and DOMINO are restricted to individuals aged between 15 and 66. The difference in means (Columns 2 and 3) are always statistically different from zero at a 5% significance level, according to two-sample *t*-tests. Exceptions only include the binary indicators for New South Wales and the Northern Territory. Census data taken on 9 August 2016. Statistics for the education levels and for marital status are calculated as a fraction of non-missing observations. Missing rates for education in Census and DOMINO are 12% and 53%, respectively. Missing rates for marital status in Census and DOMINO are 11% and 0.1%, respectively. All variables are described in Appendix A in Data S1.

### Dependent (outcome) variable

We calculate the intensity of welfare receipt as the proportion of time an individual received an income support payment during the four-year period spanning 2015 to 2018: that is, the number of days they received an income support payment from 1 January 2015 to 31 December 2018 divided by the total number of days over this period. We have chosen this 4-year period over which to calculate the duration of welfare receipt because it is long enough to identify long-term welfare dependence but also short enough so that we do not obscure the individual's needs at different stages of the life cycle.

Our measure of welfare receipt intensity is consistent with how welfare reliance has been conceptualised by previous studies and in government reviews such as the most recent McClure review (Reference Group on Welfare Reform, 2015). It also captures a more severe form of economic disadvantage than measures that only consider the receipt of income support at one point in time. This is important since receipt of welfare payments is not the same thing as being dependent on welfare – a point that has been well emphasised in the literature (Penman, 2006).

Long-term welfare receipt is a prevalent issue in Australia. For example, in 2018, almost 3 in 4 income support recipients aged 18–64 had been on a payment for 2 years or more and among those receiving income support payments in 2009, more than half (56%) were receiving payments 9 years later in 2018 (AIHW, 2019a). This issue is also increasing in importance. In 2018, 24.5% of unemployed people aged 15 and over had been looking for work for more than a year (annual average), increasing from 14.8% in 2009 (ABS, 2019a).

Figure 1 plots the distribution of income support intensity. In Panel a of Figure 1, focusing on everyone in our sample, we see spikes in the proportion of people who never received income support (32.3% of the sample), and those who received income support consistently over the period 2015-18 (36.7% of the sample).[4] In Panel b of Figure 1, we display the distribution of income support intensity once we exclude the two extreme cases. Here, we see a more uniform distribution across the spectrum of intensities with a slightly stronger concentration at the higher end.

As an extension, we also predict the outcome of long-term unemployment benefit receipt. This proxies for long-term unemployment, which is particularly relevant given the recent recession induced by the COVID-19 pandemic and the attendant high rate of unemployment. Furthermore, focusing on unemployment benefits is useful because long-term receipt of this payment indicates a failure to achieve its primary purpose: to act as a temporary payment while people transition from unemployment to employment. Moreover, as previously mentioned, our ML models can potentially fulfil a useful governance and auditing role by ensuring that individuals who are most at risk of being long-term unemployed are not unduly ignored through 'cream-skimming' when third-party employment service providers make resource allocation decisions.

Figure 2 plots the distribution of unemployment benefit intensity. Similar to the pattern of any type of income support receipt, we see spikes in the proportion of people who

[4] This type of bi-model pattern is common across the entire period between 2000 and 2019.

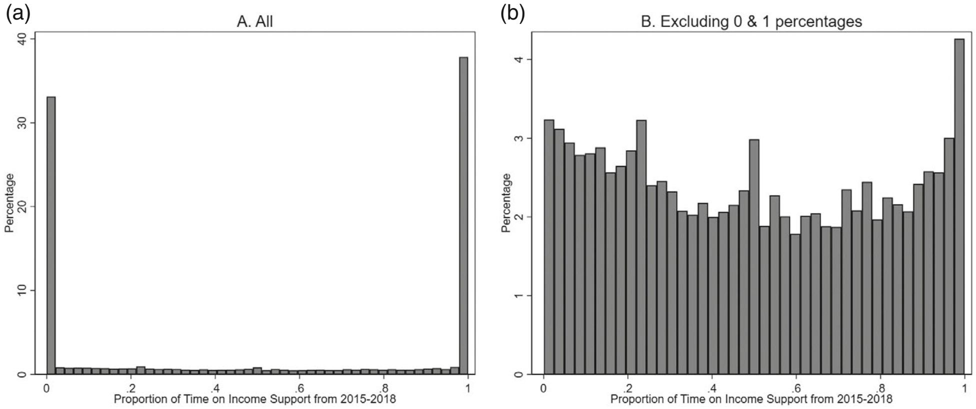

Figure 1. Proportion of time on income support between 2015 and 2018. *Notes*: These plots show the distribution for the proportion of time individuals received an income support payment between 2015 and 2018. Individuals with no income support or always on income support are excluded in Panel b

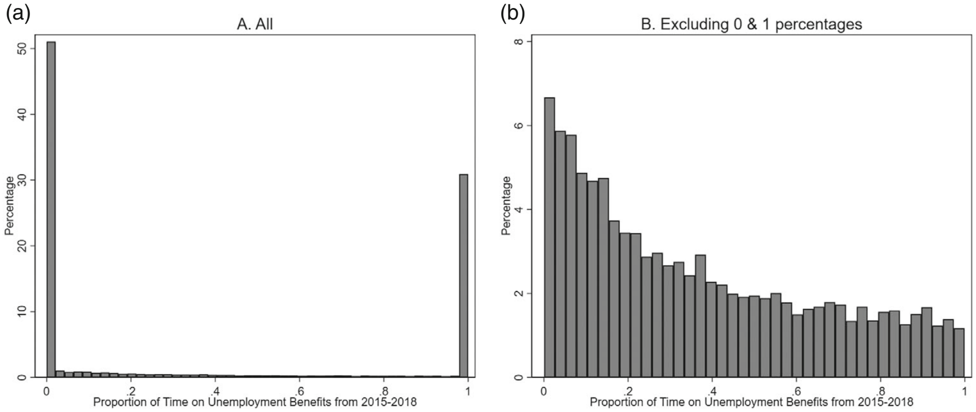

Figure 2. Proportion of time on unemployment benefits between 2015 and 2018. *Notes*: These plots show the distribution for the proportion of time individuals received unemployment benefits between 2015 and 2018. Individuals with no unemployment benefits or always on unemployment benefits are excluded in Panel b

never received unemployment benefits (50.5% of the sample), and those who received unemployment benefits consistently over the period 2015–18 (30.3% of the sample). This is illustrated in Panel a of Figure 2. Once we exclude the two extreme cases (Panel b of Figure 2), we see a higher concentration of intensities at the lower end.

An advantage of using these continuous outcome variables rather than binary variables indicating whether an individual was on income support for more than a certain time is that it provides more information and flexibility. Indeed, rather than making an arbitrary initial decision of which cut-off to use – for example, whether to focus on predicting those on income support for 50% rather than 80% of the time period – we allow this decision to be done afterwards. For instance, government officials may decide to target individuals predicted to be on income support for 90% of the time period rather than 80% or 70% if limited resources do not allow a more extensive targeting.

Alternatively, based on supplementary causal analysis, caseworkers may believe that a certain intervention may not be effective for individuals always on income support, and thus focus on those predicted to be on income support between, say, 50% and 75% of the time period. Policymakers may also decide to prioritize certain subgroups, and thus use different thresholds for different sub-populations (Kleinberg *et al.*, 2018). Furthermore, another advantage of using a continuous variable is that one can use a standard benchmark such as the Mean Squared Error (MSE). Researchers predicting binary outcomes have to choose between a large number of goodness-of-fit criteria (e.g. accuracy rate, pseudo-$R^2$, McFadden-$R^2$, recall rate, AUC), often leading to arbitrary decisions (Sansone, 2019).

### Independent (input) variables

The DOMINO data capture a wide range of information on individuals. Based on data observed in 2014 (the base year), we include information on: demographics (sex, age, country of birth, and Indigenous status); household structure (parent status, number of children, and ages of children), government benefit receipt history (by benefit type); health issues experienced by the individual or by family members (according to receipt of disability benefits or carer benefits); personal relationships (partnership status and marital status, as well as the duration and instability of relationship status); employment and underemployment (zero-hour contracts); work instability; location and residential mobility; housing; education; income and wealth.

A key benefit of the high frequency DOMINO data is the ability to model the dynamics of economic behaviour. For example, we include variables that capture the instability, variability and intensity (number or duration) over time of income support history and employment. Examples include: exits and entries onto the income support system, annual fluctuations in the amount and duration of benefits, wages and earnings, hours of employment, number of jobs (including simultaneously held jobs) and changes in the working arrangements (day-of-week and hours per week) both within the same employer and across employers.

We also proxy for the individual's non-cognitive ability or behavioural tendencies, such as risk preferences or forward-looking behaviour, by using an array of (over 60) measures of intensity and volatility of past employment spells and income support history (see Appendix A). Such labour market history variables have been shown to act as good proxies for behavioural traits in predicting unemployment and reemployment success, as well as reflecting motivation and an individual's private information about skills, thus absorbing the variation from traits such as years of education, self-reported job search and conscientiousness (Van Landeghem *et al.*, 2021; Raveendran, Puranam, and Warglien, 2022). In addition, we include measures such as whether the individual was ever sanctioned or had their benefit cancelled because of non-compliance reasons. Indeed, one interpretation of being sanctioned may be disorganisation or delays (Banerjee and Duflo, 2014) on the part of the registrant (to fulfil mutual obligations or activity test requirements, for example). However, we recognize that these measures are likely to reflect other factors and constraints as well (Klapdor, 2013).

## IV. Methods

### Simple and heuristic models

We begin by providing a benchmark for the ML performance and to proxy current methods used by practitioners when predicting individuals at-risk of welfare dependency. To do this, we estimate simple

ordinary least square (OLS) models predicting the proportion of time an individual spends on income support between 2015 and 2018, as reported in Table 2 and Figure 3. The most basic benchmark is a regression containing only a constant term (Model 1). Building on this, we evaluate the predictive power of demographic characteristics: sex (Model 2), education (Model 3) and age (Model 4). We then consider the predictive power of income support history – that is, a series of binary variables indicating whether an individual had ever received an income support payment at any time in 2014, separately for different types of income support payments – in Model 5.

Another natural benchmark is to estimate a heuristic model, which includes variables that the past literature has identified as key drivers of welfare dependency such as sex, age, education level, parent status, migration status, ethnicity, marital status, state of residence and unemployment status (Model 6). We further enhance this model by adding the income support history from Model 5 (Model 7). In addition, given the emphasis in the past literature on local neighbourhood effects (Chetty *et al.*, 2018), we include detailed geographical information in Model 8.

Finally, we test the potential predictive gain of ML models over the profiling indicators currently used by the Australian government to prevent welfare dependency (Model 9). As outlined in Section II, these are indicators for the three groups with the largest estimated lifetime welfare costs: young carers aged 24 or under, students receiving a particular income support payment, and young parents aged 18 or under.

### ML approach

The heuristic models rely on a small set of variables and on methods like simple aggregation for data reduction. Yet, in practice, simple models are unlikely to represent the complex processes underlying welfare dependency. Fortunately, our baseline dataset enables greater flexibility because it contains a rich set of possible predictors. We include a wide array of these predictors and different functional forms for each predictor. For example, we include a full set of indicator variables for a recipient's country of birth, their marital status, employment and education status, along with other nonlinear expressions (and complex interactions) between these variables (as described in detail in the Data S1).

Following Mullainathan and Spiess (2017), we do not drop redundant or aggregated variables since they could be useful to obtain better predictions with less complexity. For example, we include categorical variables recording the number of children in the household while also including an extensive set of binary indicators indicating family size and fertility progression. The net result of this is that we have approximately 1,800 possible predictors. Moreover, if we start considering interaction terms between some of these predictors, the number of variables can reach or even exceed the number of individuals in the data. In a standard OLS regression, this would not be possible.

TABLE 2

*Prediction (out-of-sample) performance for time on income support*

| | | | | *MSE* | | |
|---|---|---|---|---|---|---|
| | | *Model* | *Predictors* | *Mean* | *CI* | *R-squared* |
| Simple models | 1 | OLS | Constant | 0.203 | [0.202; 0.205] | 0.0% |
| | 2 | OLS | Sex | 0.202 | [0.200; 0.203] | 0.9% |
| | 3 | OLS | Education | 0.197 | [0.195; 0.199] | 3.0% |
| | 4 | OLS | Age | 0.168 | [0.165; 0.171] | 17.4% |
| | 5 | OLS | Income support history | 0.083 | [0.080; 0.085] | 59.5% |
| Economist models | 6 | OLS | Heuristic | 0.141 | [0.138; 0.144] | 30.6% |
| | 7 | OLS | Heuristic + Income support history | 0.077 | [0.074; 0.079] | 62.3% |
| | 8 | OLS | Model 7 + Location | 0.077 | [0.074; 0.079] | 62.5% |
| Actuarial | 9 | OLS | Carer + Student + Parent | 0.190 | [0.188; 0.192] | 6.5% |
| Machine learning | 10 | LASSO | All baseline inputs | 0.052 | [0.050; 0.054] | 74.7% |
| | 11 | SVR | All baseline inputs | 0.050 | [0.048; 0.052] | 75.4% |
| | 12 | Boosting | All baseline inputs | 0.050 | [0.048; 0.053] | 75.5% |
| | 13 | Ensemble | All baseline inputs | 0.048 | [0.046; 0.050] | 76.3% |

*Notes*: The outcome variable is the proportion of time an individual received an income support payment in 2015–18. Each row is a different model. MSE is the out-of-sample MSE computed using the 20% hold-out sample (Step 3 as described in Appendix B.1 in Data S1.). The numbers in brackets are bootstrapped 95% confidence intervals for hold-out prediction performance (Step 4 in Appendix B.1 in Data S1.). *R*-squared is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. Model 6 inputs (measured in 2014, corresponding missing indicators added whenever necessary): sex, age, migration status, ethnicity, parent status, marital status, state of residence, education level, unemployment status. Location data in Model 8 include SA3 region data: there are 329 unique geographic regions. The actuarial Model 9 identifies three groups with the largest estimated lifetime welfare costs: young carers aged 24 or under ('Carer'); students receiving a particular income support payment ('student'); and young parents aged 18 or under ('parent'). All variables are described in Appendix A in Data S1. The sample size of the hold-out sample is 10,123.

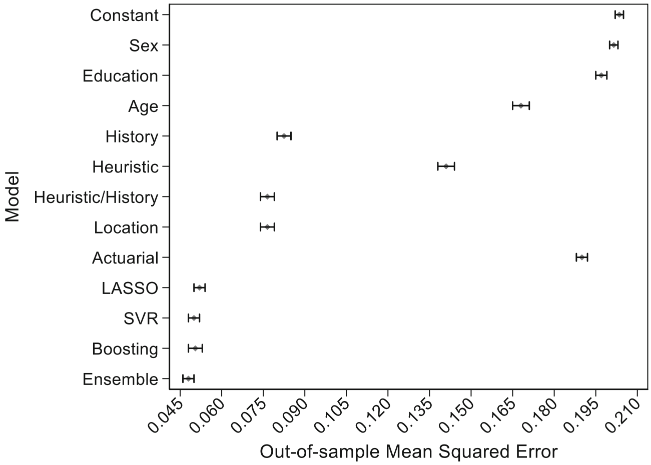

Figure 3. MSE prediction (out-of-sample) for time on income support. *Notes*: The outcome variable is the proportion of time an individual received an income support payment in 2015–18. Each row is a different model. Each diamond is the average out-of-sample MSE for a given model, plotted within the corresponding 95% confidence intervals. See also notes in Table 2. The ordering of the models is the same as in Table 2

By contrast, ML efficiently handles high-dimensional data. This allows us to identify previously undetected relationships between variables. A common issue with adding complexity to a model (in the form of more variables, more interactions, or more flexible functional forms), however, is overfitting. This issue is evident when an estimated model tightly fits a training sample but poorly fits new samples. The problem of overfitting cannot be easily solved by estimating simple models (e.g. with fewer variables subjectively selected) because this may result in models that underfit the data and exclude powerful predictors. For instance, if the true relation between $y$ and $x$ is quadratic, a linear model would be an under-fit (high bias), while estimating a fourth degree polynomial would lead to over-fitting (high variance). ML addresses the bias-variance trade-off by carefully reducing the number of variables, as well as through a process called regularization. The latter keeps all the variables but reduces the magnitude of each coefficient. This approach works well when there are numerous variables that contribute to predicting the outcome in a statistically significant, albeit economically modest, fashion. We use three different ML algorithms which employ one or both of these methods for dimension reduction.

Estimating a range of ML models (as we do) is advisable because each have their benefits and drawbacks. As noted by Athey and Imbens (2019), there are no formal results that show that one ML approach is better than another. Therefore, the choice of which algorithms to use can often be rather arbitrary. We choose to estimate different classes of ML models, including LASSO, Support Vector Regression (SVR) and Boosting (reported in Table 2 Models 10–12 and Figure 3). This allows us to evaluate algorithms which have different levels of flexibility and interpretability. LASSO is straightforward to explain given its similarity with OLS and can be easily interpreted. SVR offers extremely flexible functional forms. Likewise, Boosting is a very flexible tree-based algorithm which can account for a large number of possible interactions among the inputs. Indeed, in this specific application we allow up to six-way interactions between input variables in our Boosting algorithm. We briefly summarise each of these algorithms in Section B of the Data S1. A more comprehensive review of the tools available to practitioners is provided by, among others, Hastie, Tibshirani, and Friedman (2009), Mullainathan and Spiess (2017) and Athey and Imbens (2019).

We follow the recommendation of Mullainathan and Spiess (2017) and split the data into two sub-samples. A training sample (80% of the data) is used to calibrate and estimate the algorithm under each of the ML methods. Out-of-sample performance is reported using the hold-out sample (the remaining 20% of the data). A detailed step-by-step description of our calibration procedure is available in Section B of the Data S1.

In many cases, a single algorithm does not perform as well as a combination of methods. Different ML algorithms may capture different features of the data. Therefore, combining these algorithms – as pioneered by Bates and Granger (1969) – may potentially lead to a superior performance (Athey and Imbens, 2019). To test this hypothesis, we aggregate in Table 2 Model 13 and Figure 3 the predictions from the three main ML algorithms (LASSO from Model 10, SVR from Model 11, and Boosting from Model 12) using weights obtained through running a linear regression of the outcome on these predicted values, as described in Mullainathan and Spiess (2017).

## V. Results

Table 2 and Figure 3 show the out-of-sample performance of the aforementioned models when predicting the proportion of time an individual received income support between 2015 and 2018 using their available information from 2014. For each model, we report the MSE to compare the predictive accuracy of different models. The MSE is a widely accepted criterion to measure the performance of models predicting continuous variables. Given its intuitive appeal, we also report the square of the Pearson correlation coefficient, that is, the correlation between the actual and fitted dependent variable. This is equivalent to the *R*-squared in linear least squares regressions. All these goodness-of-fit measures are estimated out-of-sample using a hold-out sample. We use a separate training sample to estimate and calibrate the models. In-sample performances are reported in Table B4.

In Models 2−4, we report that demographic characteristics have limited predictive power: OLS models including only information on sex or education have an out-of-sample *R*-squared of 3% or lower (Models 2 and 3). The OLS model using age as the predictor explains a higher percentage of the variability in the outcome variable (17.4%, Model 4). This is to be expected given the age requirements in several income support programmes. Despite the improvement in predictive performance over Models 2 and 3, the *R*-squared remains low in absolute terms.

In Model 5, we show that income support history is strongly correlated with future welfare dependency. Specifically, a simple OLS model that includes binary indicators for any income support payment (separately for different types of payments) received in 2014 can explain almost 60% of the variability in the proportion of time on income support in the four subsequent years.

In Model 6, we find that the heuristic model does not improve performance in a substantive way. This is surprising as the set of predictors of demographic characteristics, location, employment and education status is commonly used by economists to explain economic behaviour. Further adding this set of predictors to Model 5 (i.e. with the income support history) only leads to a small reduction in the MSE. This suggests that our indicators of income support history are stronger predictors of future welfare receipt intensity. We obtain similar conclusions when we expand the set of controls to include detailed geographic information (Model 8). In addition, we show that the three at-risk groups identified in the ‘Try, Test and Learn’ programme have low predictive power when modelled in an OLS specification (Model 9).

Importantly, the results in Table 2 and Figure 3 also indicate that using ML to exploit a larger set of information already available in administrative data (that are accessible by caseworkers) lead to substantial performance gains (Models 10−13). Consequently, ML predictions, once integrated into the current identification approaches used in the ‘Try, Test and Learn’ programme, could act as a first step in targeting early interventions at the most at-risk individuals. These improvements are consistently achieved irrespective of the particular ML algorithm implemented. The stability of the performance across algorithms is both reassuring and surprising given the contrast in their data-fitting mechanics (in terms of the flexibility allowed for functional forms and interaction terms). Combining all three algorithms (LASSO from Model 10, SVR from Model 11, and Boosting from Model

12) in Model 13 further improves prediction performance.[5] It is important to note that we obtain this out-of-sample predictive accuracy without overfitting the data (as evident in Table B4) and by using a 1% random sample of the full set of observations. In our context, it suggests that further sample size augmentations (and the attendant computational costs) are unnecessary to properly calibrate these ML algorithms.

The Ensemble method, compared to the benchmark models (Models 1–9), performs significantly better. These performance gains are clear visually in Figure 3. Specifically, the out-of-sample MSE for the Ensemble model is less than one-fourth of the corresponding MSE of the basic benchmark model using only a constant (Model 1). Furthermore, its MSE represents a 42% reduction from the MSE of Model 5 with only the income support history. Bootstrapped confidence intervals are also tight around the mean MSE, suggesting that all comparisons in predictive performance between OLS and ML models are significantly different from zero. Similarly, the out-of-sample R-squared jumps to more than 76%, almost a 14-percentage point (or 22%) increase – again, at no extra cost – compared to the OLS model with the highest out-of-sample R-squared (Model 8). According to our back-of-the-envelope calculations of the total annual accrued welfare costs, the individuals identified in the ML model accrued an additional welfare cost of 0.99 billion AUD compared with the comparably sized (three) groups identified in the actuarial model. To give a sense of the magnitude, this represents roughly 10% of the total annual unemployment benefit expenditure (AIHW, 2019b).

Given the results in Model 5 in Table 2, agencies lacking the time or resources to develop and apply more sophisticated models could identify – with a medium level of accuracy – individuals already in the system who are at risk of relying on income support for an extended period of time using only past income support payment information. Alternatively, one could focus on the top predictors from LASSO (Model 10 in Table 2) and Boosting (Model 12 in Table 2): Table 3 lists the most powerful predictors selected by these two algorithms. While most of these variables are not surprising, a subset of them – such as income fluctuations, number of residential moves, and failure to complete means tests – is not routinely used in simple early warning systems. As shown in Models 1–3 in Table 4, using these selected inputs in an OLS model would yield high out-of-sample performance.[6] Therefore, while ML algorithms reach the lowest out-of-sample MSE, insights from these algorithms can be easily incorporated into simpler modelling frameworks such as an OLS model. This is especially useful if technical, political, or administrative barriers prevent ML algorithms from being implemented at a larger scale.

## VI. Extensions and robustness checks

### Alternative predictors, seed number, and cross-validation folds

Table 4 reports additional extensions and robustness checks. Despite documented large economic inequalities between native and non-native populations (Markham and

[5] This improvement is achieved despite the correlations of the out-of-sample predictions between the LASSO, SVR, and Boosting models being very high (between 0.97 and 0.99), suggesting that even under these conditions there are additional benefits of weighting and combining different algorithms.
[6] Table 4 shows out-of-sample predictions. We report the corresponding in-sample performance in Table B5.

TABLE 3

*Top 10 predictors selected by the LASSO and Boosting algorithms*

| *Order of importance* | *Level of importance* | *Predictors (calculated in 2014)* | *Variable name* |
|---|---|---|---|
| *LASSO* | *Post-LASSO coefficients* | | |
| 1 | −1.482 | Benefit fluctuation | p_sdpy |
| 2 | 0.423 | Total benefit received | p_totpy2014 |
| 3 | 0.345 | Benefit duration (any income support payment) | p_isdur14 |
| 4 | −0.341 | Annual employment income | p_totinc2014 |
| 5 | −0.270 | Wage rate | p_wage2014 |
| 6 | 0.236 | Number of residential moves | p_totmoves |
| 7 | −0.219 | Number of sanctions | p_numsus2014 |
| 8 | 0.216 | Age: below 25 (on 1 January 2014) | p_aged1 |
| 9 | −0.214 | Maximum number of simultaneous jobs held | p_maxsimjob2014 |
| 10 | 0.214 | Received benefit in December | p_Dec2014 |
| Boosting | Influence Parameters | | |
| 1 | 69.538 | Benefit duration (any income support payment) | p_isdur14 |
| 2 | 11.584 | Number of jobs (missing) | p_numjob2014miss |
| 3 | 4.313 | Received benefit in quarter four | p_qr42014 |
| 4 | 3.788 | Benefit terminated for non-fulfilment of activity test | p_evsannb2014 |
| 5 | 3.267 | Benefit duration on non-main income support benefits | p_othdur2014 |
| 6 | 1.514 | Income fluctuation | p_sdinc |
| 7 | 0.605 | Age: 59–64 (on 1 January 2014) | p_agecatd9 |
| 8 | 0.584 | Ever on age pension payment | p_evage14 |
| 9 | 0.533 | Benefit duration (disability support pension) | p_dspdur2014 |
| 10 | 0.320 | Annual hours | p_tothr2014 |

*Notes*: This table lists the top 10 inputs selected by LASSO (Model 10 in Table 2) and Boosting (Model 12 in Table 2) when predicting the proportion of time an individual received an income support payment in 2015–2018. All listed variables were measured in 2014. LASSO selected 323 variables and Boosting selected 72 variables. The missing indicator selected by Boosting – 'Number of jobs (missing)' – is perfectly collinear with the variable 'Ever on Income Support' (p_evis14). In LASSO, the reported coefficients are computed by taking the selected variables and estimating an OLS regression with them. In Boosting, the influence of an input depends on the number of times a variable is chosen across all iterations (trees) and its overall contribution to the log-likelihood function; such values are then standardised to sum up to 100. All variables are described in Appendix A.

Biddle, 2018), Indigenous status has limited predicted power in this context (Model 4). Similar conclusions are reached when focusing on country of birth (Model 5). In Model 6, to investigate how much income support dependency is affected by local factors, we further extend our set of detailed geographical inputs to include almost 2,500 postal codes, but we do not improve substantially upon the model that includes only information on income support history (Model 5 in Table 2). Including indicators of parental welfare receipt – a factor that has been shown to increase the likelihood of one's own welfare receipt (Dahl *et al.*, 2014; Dahl and Gielen, 2021; Cobb-Clark *et al.*, 2022) – to the heuristic models has only a marginal effect on out-of-sample performance (Models 7 and 8).

It is worth emphasizing that, as discussed in Section B of Data S1, Boosting weights and combines different predictions from several (tree-based) classifiers, and at each iteration it uses only a random subset of the training data to build such trees. Therefore, this algorithm is more robust to outliers than LASSO, and the resulting list of top predictors reported in Table 3 is less likely to be influenced by specific observations or values in the training sample. To verify that our LASSO performance is not driven by a few outliers, we can re-calibrate the LASSO algorithm using a different seed number. As shown in Model 9 of Table 4, the out-of-sample performance is almost identical to the one reported in Model 10 of Table 2. In addition, the list of top predictors remains very similar to the one reported in Table 3. For instance, government benefit fluctuation in 2014, total amount of government benefit received in 2014, annual employment income in 2014 and wage rate in 2014 are selected again among the top five predictors. Proportion of time spent on income support in 2014 is selected with the new seed among the top five predictors instead of total number of days receiving benefit in 2014, but the two variables are highly correlated, so this is not surprising. Overall, we find no evidence that the main results from LASSO in Tables 2 and 3 are sensitive to outliers.

Similarly, we show that five folds are enough to properly calibrate the algorithms with cross-validation. First, in-sample (Table B4) and out-of-sample performance (Table 2) are very close to each other. Second, as shown in Model 10 in Table 4, using a 10-fold

TABLE 4

*Prediction (out-of-sample) performance for time on income support. Extensions*

| | | | *MSE* | | |
|---|---|---|---|---|---|
| | *Model* | *Predictors* | *Mean* | *CI* | *R-squared* |
| 1 | OLS | Top 10 predictors from the LASSO model | 0.073 | [0.057; 0.061] | 64.1% |
| 2 | OLS | Top 10 predictors from the Boosting model | 0.056 | [0.054; 0.058] | 72.5% |
| 3 | OLS | Union of the top predictors from Models 1 and 2 | 0.056 | [0.054; 0.058] | 72.6% |
| 4 | OLS | Indigenous | 0.201 | [0.199; 0.203] | 1.3% |
| 5 | OLS | Country of birth | 0.198 | [0.196; 0.200] | 2.6% |
| 6 | OLS | Heuristic + Income support history + ZIP code | 0.080 | [0.077; 0.082] | 60.8% |
| 7 | OLS | Heuristic + Parental welfare receipt | 0.140 | [0.137; 0.143] | 31.3% |
| 8 | OLS | Heuristic + Income support history + Parental welfare receipt | 0.080 | [0.074; 0.079] | 62.6% |
| 9 | LASSO | Full model (different seed number) | 0.053 | [0.051; 0.054] | 73.5% |
| 10 | LASSO | Full model (10-fold cross-validation) | 0.052 | [0.050; 0.054] | 74.7% |

*Notes*: The outcome variable is the proportion of time an individual received an income support payment in 2015–18. Each row is a different model. See also notes in Table 2. Location data in Model 6 include a total of 2,456 ZIP codes (i.e. postcodes). Model 7 includes a separate set of binary variables recording the income support receipt history of the primary female carer, the primary male carer and any carer. Income support payments include: any type of income support; disability payment; unemployment benefit; carer payment; parenting payment; partner payment; age pension; financial difficulty payments.

cross-validation procedure instead of a five-fold cross-validation to calibrate the LASSO algorithm leads to almost identical predictions.

### Different functional forms and outcome variables

Table 5 presents additional robustness checks exploring different functional forms and ways to define the outcome variable.[7] Models 1–3 use the same variables used in the heuristic Model 6 in Table 2, but as inputs in a Fractional Outcome Regression (FOR) – Probit specification (to account for the clumping in the outcome variable at 0 and 1, as shown in Figure 1), and as inputs in a Boosting algorithm and a LASSO algorithm rather than an OLS model. The out-of-sample $R$-squared increases from 30.6% to 31.2% for the FOR; and it increases to 37.3% for the Boosting algorithm. This suggests that interaction terms and nonlinearities do play a role in this context, but they are not enough to reach the same performance as the OLS model with the income support history variables or the ML algorithms with the full set of inputs. Furthermore, as shown in Model 3 in Table 5, LASSO with the same inputs as the heuristic model (Model 6 in Table 2) does not improve upon the OLS model.[8]

Model 4 in Table 5 presents results for a LASSO model where we manually add interaction terms to the main LASSO model (from Model 10 in Table 2).[9] The added gain of including additional interaction terms (two-way interactions between the top 20 predictors from the LASSO in Model 10 in Table 2) is minor.[10] For example, the

TABLE 5

*Prediction (out-of-sample) performance for time on income support. Different functional forms*

| | | | *MSE* | | |
|---|---|---|---|---|---|
| | *Model* | *Predictors* | *Mean* | *C.I.* | *R-squared* |
| 1 | FOR | Heuristic | 0.140 | [0.137; 0.143] | 31.2% |
| 2 | Boosting | Heuristic | 0.128 | [0.124; 0.131] | 37.3% |
| 3 | LASSO | Heuristic | 0.141 | [0.138; 0.144] | 30.6% |
| 4 | LASSO | All baseline inputs (plus interactions) | 0.049 | [0.047; 0.051] | 76.0% |
| | | | *Accuracy* | | |
| | | | *Mean* | *C.I.* | *Recall* |
| 5 | Logit binary | Heuristic | 75.9% | [75.1%, 76.7%] | 53.9% |
| 6 | Logit binary | Heuristic + Income support history | 78.1% | [77.2%; 79.0%] | 74.2% |
| 7 | LASSO binary | All baseline inputs | 88.8% | [88.2%; 89.4%] | 86.7% |

*Notes*: The outcome variable in Models 1–4 is the proportion of time an individual received an income support payment in 2015–18. The outcome variable in Models 5–7 is a binary variable equal to one if an individual always received income support payments in 2015–18, zero otherwise. Each row is a different model. See also notes in Table 2. FOR stands for Fractional Outcome Regression (Probit).

[7] Table 5 shows out-of-sample predictions. We report the corresponding in-sample performance in Table B6.

[8] This can be explained by the fact that LASSO only penalized one variable in this simulation: i.e., it chose 34 of the 35 variables in the heuristic model.

[9] We restricted the two-way interactions to the top 20 predictors for computational reasons.

[10] Table B7 lists the top 10 predictors from the LASSO with interactions model (Model 4 Table 5). Note there is a high correlation in the base predictors chosen in this model with those chosen in Model 10 Table 2, suggesting stability in the importance of these key predictors.

out-of-sample *R*-squared in Model 4 in Table 5 is only 1.3 percentage points higher than the LASSO in Model 10 Table 2 where we do not include any interaction terms. The predictive gains of Model 4 Table 5 over and above the SVR (Model 11 in Table 2) and Boosting (Model 12 in Table 2) are even smaller, pointing to the fact that these latter two algorithms automatically allow for flexible functional forms.

As already discussed, our analysis focuses on predicting the proportion of time an individual is receiving income support rather than a binary variable recording whether an individual is always on income support in order to achieve greater flexibility for practitioners and policymakers. This choice of a continuous outcome variable also allows us to use the MSE as the main goodness-of-fit criterion. An alternative approach to the FOR to account for the polarized distribution of this outcome variable is to generate a new binary outcome variable taking value one if an individual was always on income support 100% of the time between 2015 and 2018, zero otherwise. In other words, this variable is equal to zero for individuals who did not receive any income support payments between 2015 and 2018, as well as for those who received payments only occasionally during the same time period.

As shown in Model 7 in Table 5, LASSO correctly predicts whether an individual was always on income support (correct 1) or only occasionally/never on income support (correct 0) 88.8% of times. This is substantially higher than the 75.9% accuracy rate achieved by the heuristic model (Model 5 in Table 5, same inputs as Model 6 in Table 2) and the 78.1% achieved by the specification including both the heuristic inputs and the information on past income support payments (Model 6 in Table 5, same inputs as Model 7 in Table 2). This is true even if both Model 5 and Model 6 in Table 5 use a Logit instead of an OLS regression to better account for the binary nature of the outcome variable. Similarly, the recall rate is between 12 and 33 percentage points higher in Model 7 than Models 6 and 5, respectively: among the individuals always on income support, LASSO correctly identifies almost 87% of them. Therefore, ML is able to accurately predict whether an individual is at risk of permanently rely on income support.

## Varying sample and external validity

Table 6 reports additional sensitivity tests varying the sample used to calibrate our algorithms.[11] Building on the arguments used to motivate the models in Table 5, we exclude those who were on income support 100% of the time from 2011 to 2014 in Models 1–3. Comparing Model 3 with that of Model 1 (heuristic model as in Model 6 Table 2) and Model 2 (same predictors as the heuristic model plus income support history as in Model 7 Table 2), we see again that the LASSO algorithm substantially improves on the predictive performance of the OLS models.

Individuals with a historical pattern of high reliance on income support may be more difficult to mobilise into employment. This can be because of the potential scarring effects of long welfare receipt durations and/or because the very factors that lead these individuals to rely on welfare for long durations also prohibit them from finding employment. For example, approximately 61% of individuals who were on income support over the entire

[11] Table 6 shows out-of-sample predictions. We report the corresponding in-sample performance in Table B8.

TABLE 6

*Prediction (out-of-sample) performance for time on income support. Varying sample*

| | | | *MSE* | | |
|---|---|---|---|---|---|
| | *Model* | *Predictors* | *Mean* | *C.I.* | *R-squared* |
| 1 | OLS | Heuristic (excluding those 100% on welfare 2011–14) | 0.119 | [0.115; 0.123] | 32.4% |
| 2 | OLS | Heuristic + Income support history (excluding those 100% on welfare 2011–14) | 0.082 | [0.079; 0.085] | 53.6% |
| 3 | LASSO | Full model (excluding those 100% on welfare 2011–14) | 0.059 | [0.056; 0.061] | 66.8% |
| 4 | LASSO space | All baseline inputs | 0.054 | [0.052; 0.055] | 73.4% |
| 5 | OLS time | Heuristic | 0.132 | [0.131; 0.134] | 42.3% |
| 6 | OLS time | Heuristic + Income support history | 0.112 | [0.110; 0.113] | 52.2% |
| 7 | LASSO time | All baseline inputs | 0.098 | [0.096; 0.100] | 58.6% |

*Notes*: Unless otherwise specified, the outcome variable is the proportion of time an individual received an income support payment in 2015–18. Each row is a different model. See also notes in Table 2. Models 1–3 exclude those who are on income support for 100% of time from 2011 to 2014, inclusive. The joint sample size of the hold-out and training samples in Models 1–3 is 36,358. This is the 1% random sample after the deletion of those on income support for 100% of time between 2011 and 2014. Out-of-sample predictions in Model 4 have been obtained using individuals living in Victoria or the Australian Capital Territory in 2014 ($N = 13{,}105$), while the remaining 36,978 individuals have been included in the training sample. Overall, the number of observations used in Model 4 totals $N = 50{,}083$. This is slightly less than the $N = 50{,}615$ observations in other models because 532 people in the sample lived in at least two states in 2014, and we have deleted them in Model 4. Models 5–7 have been calibrated using as outcome variable the proportion of time an individual received an income support payment in 2015–16, while then predicting out-of-sample the proportion of time an individual received an income support payment in 2017–18. The sample size of the hold-out sample in Models 5–7 is 50,615.

period of time from 2011 to 2014 were either receiving disability support pensions themselves or caring for a family member with a disability (and thus receiving carers payments). In this set of sensitivity analysis, our motivation for excluding the individuals who are long-term welfare dependent from our sample in Models 1–3 in Table 6 is because conventional early intervention programmes of a labour market activation nature may be less effective when targeted at these individuals. In fact, recipients of disability support pensions and carers payments are not required to fulfil the same participation requirements as those receiving unemployment benefits. Reflecting this, the former group are not even referred onto the Job Service Australia case managers who administer labour market activation interventions only for those who are in receipt of unemployment benefits.

It is important to highlight that we nominate to include the high-reliance group in our main analysis sample because policymakers explicitly aim to assist both those who are currently long-term welfare dependent, as well as those who are at-risk of it. Identifying both these groups in our algorithms does not preclude a differentiated approach in the type of assistance or intervention programme that is targeted at them. For example, identifying those who are most likely to be long-term welfare dependent because of enduring health issues warrants the more generous programme of monetary assistance and lower participation requirements that are currently in place; whereas those who are at-risk of being long-term jobseekers may be better helped with labour market activation strategies. Furthermore, within the group of long-term jobseekers, some of them may require more intense case management (or different forms of activation strategies) than others. If individuals in the high-reliance groups receive the generic support programmes (or worse, are systematically ignored by case managers because of the aforementioned ‘cream-skimming’ issues), then the chances of mobilizing them into employment may be low. In summary, individuals who are identified as being at-risk of long-term welfare receipt are a heterogeneous group and are likely to have a different set of needs. The treatment strategies aimed at them should be tailored accordingly. We explore this in more detail when we estimate unsupervised ML models in the next section.

One additional concern is with respect to external validity: we trained our models on Australian data and for a specific time period, which may or may not limit the generalisability of our ML models. Before setting out to test this, *a priori*, we expect there to be broad interest in the Australian case study for a number of reasons. First, most advanced economies are building technological infrastructures using administrative data similar to the Australian one. Some European countries with public-health systems have even additional information such as detailed health data. Second, some low-income countries are also starting to build their data capabilities. Third, welfare payments such as unemployment or disability benefits are common in many countries, and they have similar means-testing criteria to the Australian ones. This makes our application of interest to policymakers and practitioners in other countries since it shows the potential of using ML with big administrative datasets in the context of welfare recipients. The increasing quality of administrative data in all these countries, as well as the similarities in the welfare system between Australia and other countries in Europe or the Commonwealth, suggest that our proposed approach could be valuable even when applied outside Australia.

In order to increase our external validity claims, we test whether the ML algorithms continue to achieve high performance by including individuals living in certain states in a training sample and individuals living in other states in a hold-out sample. There are significant differences in the states and territories of Australia stemming from the historical separation of these jurisdictions. Specifically, our hold-out sample includes individuals living in 2014 in Victoria or the Australian Capital Territory, while the training sample includes individuals living in all other states. As shown in Model 4 of Table 6, also in this case LASSO manages to achieve an *R*-squared of 73.4% and a low MSE comparable to the main one in Model 10 of Table 2. This performance remains superior to the ones of the heuristic and actuarial models reported in Table 2. Therefore, even if we cannot test whether our algorithm could achieve high performance also outside Australia, we can at least confirm that our findings are geographically stable within Australia. Even if each state varies in term of socio-economic characteristics, practitioners could reliably use algorithms trained in one state in another state.

We can also compare ML algorithms to other models using different time periods in the training and in the hold-out samples. This is a good test to check for the stability of the ML performance over time. In particular, we calibrate Models 5–7 in Table 6 using as outcome variable the proportion of time an individual received an income support payment in 2015–16 rather than the longer period 2015–18 exploited in Table 2. The input variables from 2014 are unaltered. We then compute the out-of-sample performance

TABLE 7

*Prediction (out-of-sample) performance for time on unemployment benefits*

| | | | | *MSE* | | |
|---|---|---|---|---|---|---|
| | | *Model* | *Predictors* | *Mean* | *C.I.* | *R-squared* |
| Simple models | 1 | OLS | Constant | 0.206 | [0.203; 0.209] | 0.0% |
| | 2 | OLS | Sex | 0.204 | [0.202; 0.207] | 0.9% |
| | 3 | OLS | Education | 0.201 | [0.198; 0.204] | 2.7% |
| | 4 | OLS | Age | 0.083 | [0.080; 0.087] | 59.6% |
| | 5 | OLS | Income support history | 0.125 | [0.122; 0.128] | 39.5% |
| Economist models | 6 | OLS | Heuristic | 0.036 | [0.034; 0.038] | 82.8% |
| | 7 | OLS | Heuristic + Income support history | 0.035 | [0.034; 0.037] | 83.1% |
| | 8 | OLS | Model 7 + Location | 0.035 | [0.033; 0.037] | 83.1% |
| Machine Learning | 9 | LASSO | All baseline inputs | 0.029 | [0.027; 0.031] | 86.2% |

*Notes*: The outcome variable is the proportion of time an individual received unemployment benefits in 2015–19. Each row is a different model. See also notes in Table 2.

reported in Table 6 by predicting the proportion of time an individual received an income support payment in 2017–18. This is a more demanding task than the in-sample split done in Table 2, but it simulates a realistic potential application of our algorithms: practitioners could use data on past welfare transfers to predict future reliance on income support payments. Remarkably, the out-of-sample ML performance remains superior to those of the heuristic models: the MSE for the LASSO (Model 7) is lower than the MSE for the heuristic model (Model 5, same inputs as Model 6 in Table 2), even when augmented with information on past income support payments (Model 6, same inputs as Model 7 in Table 2). The 95% confidence intervals do not overlap, and the *R*-squared is higher for LASSO. Therefore, even if the gains from using ML are lower in absolute terms in this simulation than in Table 2, they remain substantially and economically meaningful.

### Long-term unemployment benefits

Finally, in Table 7, we estimate another set of results using the outcome of long-term unemployment benefit receipt rather than the more general long-term welfare receipt analysed so far.[12] The ML models perform slightly better than the benchmark models. For example, the out-of-sample *R*-squared increases from 83.1% to 86.2% from Model 8 to Model 9, and the bootstrapped 95% confidence intervals for the MSE do not overlap.

It is important to note that while the gain from ML seems to be larger when the algorithms are applied to the whole sample of welfare recipients in Table 2 (and even if the predictive gain from LASSO in Table 7 is only 3 percentage points) the cost of calibrating an ML model on existing datasets is negligible. This makes it worthwhile to estimate ML models over simpler OLS models also in this context.

[12] Table 7 shows out-of-sample predictions. We report the corresponding in-sample performance in Table B9.

## VII. Unsupervised ML

As clear from our discussion in the previous sections, individuals who receive income support intensively are unlikely to be a homogenous group. There may be more narrowly defined populations because of two factors. First, the Australian welfare system is structured to provide different payments to address distinct needs in the population (e.g. disability benefits are provided to those who have severe health issues; unemployment benefits are provided to those who are unemployed; and parenting benefits are provided to primary carers of young children). Reflecting this, different payments have varying eligibility criteria, as well as different levels of generosity. This means individuals receiving different income support payments are likely to be receiving payments for varying lengths of time simply because of their pre-existing characteristics and conditions. Second, among those receiving the same payments, the reasons for which an individual is receiving a payment can differ. For example, unemployment payment recipients comprise of individuals who are transitioning from study to employment and those who are discouraged jobseekers.

Thus, if caseworkers and welfare administrators want programmes to address individual needs and circumstances, and to effectively make an efficiency-fairness trade-off, they may want to design different treatments. Unsupervised ML can be a useful first step in undertaking this task because it identifies the distinct clusters within the population of long-term welfare recipients. In addition, it can partition clusters according to complex interactions between different variables. This is important because individuals may be at risk of long-term welfare receipt due to a complex set of circumstances, not just due to one dimension alone, such as based on their age, ethnicity or gender. In other words, this section acknowledges that long-term dependency on income support is a multidimensional issue: different factors may underlie prolonged receipt patterns. This is similar to the multidimensional approach advocated in poverty studies (Alkire and Foster, 2011). This section thus shows how individuals predicted to be at risk can be divided into different subgroups using unsupervised ML.

So far, we have focused on predictive models, that is, on supervised ML. Indeed, supervised algorithms are provided with a certain number of 'right' answers, that is, actual *y* associated with certain *Xs*, and are asked to produce other correct answers, that is, to predict new *y* given other combinations of *Xs*. On the other hand, unsupervised learning algorithms derive a structure for the data without necessarily knowing the relationship between *x* and *y*.

The starting point is the prediction obtained using the LASSO algorithm in Table 2 (Model 10). We have then used the top variables selected by LASSO (as reported in Table 3) combined with the indicators from the heuristic model and the income support history variables in Table 2 (Model 7) to divide the individuals predicted to be at risk into different groups by means of a hierarchical clustering algorithm (Stata, 2019). More formally, cluster analysis specifies whether the joint density of the observable variables *X* can be represented by a mixture of simpler densities representing distinct groups of observations. Conceptually, the hierarchical clustering algorithm can be summarized as follows: initially there are *n* distinct groups, one for each observation; in the next step, the two closest observations are merged into one group, thus resulting in $n - 1$ groups;

TABLE 8

*Clustering: Key characteristics. At-risk individuals are not a homogeneous group*

| *Predictors* | *Group 1* | *Group 2* | *Group 3* | *Group 4* | *No benefit* |
|---|---|---|---|---|---|
| Simultaneous jobs worked (number) | 0.013 | 0.013 | 0.025 | 0.015 | 0.004 |
| | (0.029) | (0.03) | (0.037) | (0.039) | (0.022) |
| Immigrant | 0.252 | 0.314 | 0.194 | 0.200 | 0.368 |
| | (0.435) | (0.464) | (0.396) | (0.400) | (0.483) |
| Indigenous | 0.081 | 0.051 | 0.155 | 0.092 | 0.013 |
| | (0.274) | (0.219) | (0.362) | (0.289) | (0.112) |
| Aged below 25 | 0.116 | 0.037 | 0.355 | 0.332 | 0.070 |
| | (0.321) | (0.19) | (0.479) | (0.471) | (0.256) |
| Aged 25–29 | 0.070 | 0.036 | 0.116 | 0.166 | 0.135 |
| | (0.255) | (0.186) | (0.321) | (0.372) | (0.342) |
| Aged 30–34 | 0.085 | 0.042 | 0.093 | 0.160 | 0.256 |
| | (0.28) | (0.201) | (0.291) | (0.366) | (0.437) |
| Aged 35–39 | 0.093 | 0.081 | 0.067 | 0.075 | 0.226 |
| | (0.291) | (0.273) | (0.249) | (0.264) | (0.419) |
| Aged 40–44 | 0.081 | 0.095 | 0.069 | 0.058 | 0.154 |
| | (0.274) | (0.294) | (0.254) | (0.235) | (0.362) |
| Aged 45–49 | 0.116 | 0.099 | 0.075 | 0.045 | 0.100 |
| | (0.321) | (0.298) | (0.263) | (0.207) | (0.300) |
| Aged 50–54 | 0.120 | 0.106 | 0.086 | 0.043 | 0.046 |
| | (0.326) | (0.309) | (0.281) | (0.202) | (0.21) |
| Aged 55–59 | 0.081 | 0.122 | 0.060 | 0.049 | 0.010 |
| | (0.274) | (0.327) | (0.238) | (0.215) | (0.099) |
| Aged 60–64 | 0.105 | 0.211 | 0.078 | 0.044 | 0.003 |
| | (0.307) | (0.408) | (0.269) | (0.204) | (0.053) |
| Ever married in 2014 | 0.221 | 0.356 | 0.143 | 0.138 | 0.736 |
| | (0.416) | (0.479) | (0.35) | (0.345) | (0.441) |
| Carer payment | 0.194 | 0.108 | 0.029 | 0.104 | 0.000 |
| | (0.396) | (0.31) | (0.169) | (0.305) | (0.000) |
| Female | 0.419 | 0.574 | 0.468 | 0.704 | 0.657 |
| | (0.494) | (0.495) | (0.499) | (0.457) | (0.475) |
| Parenting payments | 0.004 | 0.077 | 0.084 | 0.424 | 0.000 |
| | (0.062) | (0.266) | (0.278) | (0.494) | (0.000) |
| Observations | 258 | 3,108 | 1,126 | 1,009 | 712 |

*Notes*: Summary statistics (mean and standard deviation) are reported for each group identified by the hierarchical clustering algorithm. Individuals who are identified as at-risk of long-term welfare receipt are divided into four groups (first four columns of results). The last column reports summary statistics for the group of individuals predicted to have no future receipt. All variables have been rescaled to be between 0 and 1. Full set of variables reported in Table B10. All variables are described in Appendix A in Data S1.

after that, the closest two groups are merged together, producing $n-2$ groups. This process continues until all the observations are merged into one large group. Therefore, the output of this algorithm is a hierarchy of groupings from one group to $n$ groups. As explained in Sansone (2019), the Caliński and Harabasz pseudo-F index and the Duda-Hart Je(2)/Je(1) index with associated pseudo-$T^2$ can help analysts to select the best number of groups.

We find that five clusters most tightly partition the feature space. Table 8 shows the summary statistics for these predicted at-risk individuals.[13] For comparison purposes, we

[13] For simplicity, only the key variables are reported in Table 8. Summary statistics for the whole set of predictors are reported in Data S1 (Table B10). We only report summary statistics for four of the five clusters since one of the

also report the summary statistics for the individuals who are predicted to never receive income support (last column). Based on these summary statistics, we see some clear similarities between the four groups of long-term welfare recipients. For example, all of these individuals had a higher incidence of tenuous employment as defined by working in more than one job at one point in time, being more likely to be Indigenous, and comprising of individuals aged at the extreme ends of the distribution (for example, below 30 and/or above 55). In addition, many of them were less likely to have been married in 2014 (our base year).

Clear differences in the background characteristics also emerge between the groups. For example, people who were more likely to receive disability benefits themselves or receive a payment on behalf of a disabled family member are overrepresented in Group 1, while Group 2 has a higher percentage of immigrants. Group 3 is overrepresented by Indigenous individuals, younger people, and those who received unemployment benefits. Finally, Group 4 is more likely to be composed of female recipients with children.

This result emphasizes the importance of not pooling together all individuals identified as at risk. In fact, placing in the same programme individuals with disabilities side by side with young Indigenous individuals or single mothers may not be the most efficient approach and may actually result in negative programme effects.

## VIII. Conclusions

In this paper, we show that ML algorithms applied to data already available to caseworkers can significantly improve predictive accuracy compared to OLS models that use a smaller set of variables or compared to currently used early warning systems. The ML predictions can be applied to potentially reduce workloads for caseworkers and to supplement their expertise in the process of identifying high-risk individuals. We find evidence that using the better-performing ML models, as opposed to the currently used early warning indicators, could lead to substantial savings in public spending because individuals identified by the former accrue on average higher welfare expenditures than those identified by the latter. We then show how unsupervised ML can be exploited to identify sub-populations among high-risk individuals.

As already mentioned in the introduction, several caveats are worth noting. First, although we have used some of the most popular ML algorithms, it is possible that more advanced algorithms or extensive grid-searches could further improve performance. In addition, these algorithms have been designed to be applied to large data sets with millions of observations and thousands of potential inputs. Even if administrative data have large numbers of observations, the number of potential inputs is often limited. For instance, we did not have any information on where welfare recipients went to college, their risk preferences, their non-cognitive skills, or the full record of their parents' education and employment history. We should also emphasise that our simulations are based on data collected during a period of stable or improving macroeconomic conditions: ML

clusters includes only five observations. For this small group, reporting their summary statistics would compromise data confidentiality rules and would fail to provide any meaningful conclusions about the groups' characteristics.

algorithms would need to be retrained using data from economic downturns in order to improve external validity claims, especially when predicting welfare dependency among individuals affected by the COVID-19 pandemic. Despite these limitations, the findings in this paper suggest that researchers and practitioners could obtain more precise predictions – at no extra costs – by exploiting the available datasets.

Relatedly, we argue in this paper that our ML results significantly improve upon simpler regression baseline models. One such baseline model includes the set of controls that economists traditionally agree are important determinants or correlates of human decision-making based on theory and/or past empirical observations. We call this model the economist's heuristic baseline model. There can be other baseline models. We display a range of models in our simulations: some of which are even simpler than the heuristic baseline model, and others that are more complex or reflecting choices done by policymakers in Australia. In all cases, our ML model results show improved performance. Simpler baseline models have the potential drawback that they may underfit the data; however, more complicated baseline models have the drawback that they may overfit the data. Finding the 'best' baseline model with which to compare our ML results is unclear, which is why we have shown a range of benchmarks. The extent to which the ML algorithms are superior relative to the baseline model depends, of course, on which baseline model is used as a comparison. Thus, the degree to which other countries would find an improvement in implementing ML algorithms, among other things, would necessarily depend on which baseline models they currently employ.

Second, some researchers and practitioners could be concerned that ML may discriminate against certain groups, or perpetuate systemic inequalities intrinsic in the data used to calibrate the algorithms (Madhusoodanan, 2021). Removing sensitive variables such as race or immigration status from the set of inputs in the algorithm may not be a solution, both because the algorithms can find alternative variables highly correlated with the removed items, and because researchers have shown that including such variables but using the predictions differently for different groups may actually increase equity and efficiency (Kleinberg *et al.*, 2018). Similarly, gender-blind systems often end up penalizing women (Hao, 2019). Furthermore, refusing *a priori* to use any algorithm that may be seen as a 'black box' and not interpretable could lead to substantial welfare losses (Holm, 2019; Babic *et al.*, 2021). In this context, the possibility that ML may predict certain subgroups such as Indigenous people to be more likely to stay on income support for an extended period of time could actually benefit these individuals if caseworkers can then target and properly support them (Desiere and Struyven, 2021).

Third, the predictions in this paper are different from treatment effect estimates that a policymaker may use to optimally allocate resources (Athey, 2017): such agents would be more interested in knowing which individuals would benefit the most from a welfare transfer. For example, in the 'Try, Test and Learn' programme in Australia, one consideration in selecting the three at-risk groups was how responsive they would be, in a causal sense, to new interventions. But even in such cases, accurately identifying a pool of high-risk individuals may be a useful first step. Furthermore, by systemising the screening process, it can potentially limit the impact of explicit and implicit biases or 'cream-skimming', as well as allow caseworkers to focus on the most critical welfare recipients. In other words, supervised ML can be used in the first stage to identify high-risk individuals, while unsupervised ML can divide these individuals into sub-groups. These findings can then complement and support separate or subsequent causal analyses to inform policymakers about the appropriate interventions for these high-risk groups. In addition, from a social welfare perspective, policymakers may prioritize allocating resources to the most vulnerable individuals even if it does not necessarily aim to maximize the marginal gains from a treatment programme.

Future research could explore how to use the predictions from these algorithms as a preliminary step in randomised control trials (Duflo, 2018). For instance, these algorithms could help researchers to identify their population of interest – for example, the subset of individuals most at-risk – and then design an RCT to find the appropriate treatment to support these welfare recipients. Chandler *et al.* (2011) provide an early example of incorporating predictive models in public policy programmes targeting specific sub-populations. In this context, ML techniques known as multi-armed or contextual bandits can regularly update algorithms with newly harvested information (on registrants) and recommend personalized treatments (Athey, 2019). Finally, future research could elucidate how ML algorithms ought to be effectively combined with human expertise, including an understanding of the contexts in which we should rely heavily on ML algorithms vs. those in which we ought to more heavily rely on humans (Raghu *et al.*, 2019), and how to address misuse of data and predictions by public officials, as well as concerns by citizens impacted by automated and opaque decisions made by algorithms (Lokshin and Umapathi, 2022). Advances in this knowledge base can ultimately improve future resource allocation decisions.

*Final Manuscript Received: December 2021*

## References

ABS. (2014). *Australian Demographic Statistics*, December 2014, ABS, Canberra.

ABS. (2019a). *Labour force, Australia, detailed: Electronic delivery*, ABS, Canberra.

ABS. (2019b). *Personal Income in Australia*, ABS, Canberra.

Ahani, N., Andersson, T., Martinello, A., Teytelboym, A. and Trapp, A. C. (2021). ‘Placement Optimization in Refugee Resettlement’, *Operations Research*, Vol. 69, pp. 1468–1486.

AIHW. (2019a). *Australia’s Welfare 2019 Data Insights*, AIHW, Canberra.

AIHW. (2019b). *Welfare Expenditure Snapshot*, AIHW, Canberra.

Aizer, A. and Currie, J. (2004). ‘Networks or neighborhoods? Correlations in the use of publicly-funded maternity care in California’, *Journal of Public Economics*, Vol. 88, pp. 2573–2585.

Alkire, S. and Foster, J. (2011). ‘Counting and multidimensional poverty measurement’, *Journal of Public Economics*, Vol. 95, pp. 476–487.

Athey, S. (2017). ‘Beyond prediction: using big data for policy problems’, *Science Magazine*, Vol. 355, pp. 483–485.

Athey, S. (2019). ‘The impact of machine learning on economics’, in Agrawal A., Gans J., and Goldfarb A. (eds), *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, Chicago, pp. 507–547.

Athey, S. and Imbens, G. W. (2019). ‘Machine learning methods that economists should know about’, *Annual Review of Economics*, Vol. 11, pp. 685–725.

Australian Senate Committee on Community Affairs. (2021). *Report on Social Services Legislation Amendment (Strengthening Income Support) Bill 2021 [Provisions]*.

Australian Treasury. (2010). *Australia's future tax system: Report to the Treasurer, Part Two Detailed Analysis*, Australian Treasury, Canberra.

Babic, B., Gerke, S., Evgeniou, T. and Cohen, I. G. (2021). 'Beware explanations from AI in health care', *Science Magazine*, Vol. 373, pp. 284–286.

Bäckman, O. and Bergmark, Å. (2011). 'Escaping welfare? Social assistance dynamics in Sweden', *Journal of European Social Policy*, Vol. 21, pp. 486–500.

Banerjee, A. V. and Duflo, E. (2014). '(Dis)Organization and success in an economics MOOC', *American Economic Review*, Vol. 104, pp. 514–518.

Barrett, G. F. (2000). 'The effect of educational attainment on welfare dependence: evidence from Canada', *Journal of Public Economics*, Vol. 77, pp. 209–232.

Bates, J. M. and Granger, C. W. J. (1969). 'The combination of forecasts', *Journal of the Operational Research Society*, Vol. 20, pp. 451–468.

Beattie, G., Laliberte, J.-W. P. and Oreopoulos, P. (2018). 'Thrivers and Divers: using non-academic measures to predict college success and failure', *Economics of Education Review*, Vol. 62, pp. 170–182.

Bhuller, M., Brinch, C. N. and Königs, S. (2017). 'Time aggregation and state dependence in welfare receipt', *The Economic Journal*, Vol. 127, pp. 1833–1873.

Bitler, M. P., Currie, J. and Scholz, J. K. (2003). 'WIC eligibility and participation', *Journal of Human Resources*, Vol. 38, pp. 1176–1179.

Bjerre-Nielsen, A., Kassarnig, V., Lassen, D. D. and Lehmann, S. (2021). 'Task-specific information outperforms surveillance-style big data in predictive analytics', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 118, e2020258118.

Bradbury, B. and Zhu, A. (2018). 'Welfare entry and exit after marital separation among Australian mothers', *Economic Record*, Vol. 94, pp. 405–423.

Card, D., Chetty, R. A. J. and Weber, A. (2007). 'The spike at benefit exhaustion: Leaving the unemployment system or starting a new job?', *American Economic Review*, Vol. 97, pp. 113–118.

Card, D., Johnston, A., Leung, P., Mas, A. and Pei, Z. (2015). 'The effect of unemployment benefits on the duration of unemployment insurance receipt: New evidence from a regression kink design in Missouri, 2003-2013', *American Economic Review*, Vol. 105, pp. 126–130.

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*, Vol. 16, pp. 894–931.

Chandler, B. D., Levitt, S. D. and List, J. A. (2011). 'Predicting and preventing shootings among at-risk youth', *American Economic Review: Papers & Proceedings*, Vol. 101, pp. 288–292.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., and Porter, S. R. (2018). *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility*, NBER Working Paper Series No. 25147), 1–93.

Cobb-Clark, D. A., Dahmann, S. C., Salamanca, N. and Zhu, A. (2022). 'Intergenerational disadvantage: Learning about equal opportunity from social assistance receipt', *Labour Economics*, Vol. 79, 102276.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R. and Putnam-Hornstein, E. (2017). 'Risk assessment and decision making in child protective services: Predictive risk modeling in context', *Children and Youth Services Review*, Vol. 79, pp. 291–298.

Currie, J. M. and Grogger, J. (2001). 'Explaining recent declines in food stamp program participation', *Brookings-Wharton Papers on Urban Affairs*, Vol. 1, pp. 203–244.

Dahl, G. B. and Gielen, A. C. (2021). 'Intergenerational spillover in disability insurance', *American Economic Journal: Applied Economics*, Vol. 13, pp. 116–150.

Dahl, G. B., Kostøl, A. R. and Mogstad, M. (2014). 'Family welfare cultures', *Quarterly Journal of Economics*, Vol. 129, pp. 1711–1752.

Davidson, P. (2019). *Is the Job Services Australia Model 'made for measure' for Disadvantaged Jobseekers?* Centre for Public Policy Employment Services for the Future Conference At: University of Melbourne.

Department of Employment, Skills, Small and Family Business. (2020). *The Evaluation of Job Services Australia 2012–2015*, Department of Employment, Skills, Small and Family Business, PricewaterhouseCoopers, Canberra.

Department of Social Services. (2018). *30 June 2017 Valuation Report*.

Desiere, S. and Struyven, L. (2021). 'Using Artificial intelligence to classify jobseekers: the accuracy-equity trade-off', *Journal of Social Policy*, Vol. 50, pp. 367–385.

Duflo, E. (2018). *Machinistas Meet Randomistas: Useful ML Tools for Empirical Researchers*, NBER Summer Institute, Cambridge.

Feldstein, M. (2005). 'Rethinking social insurance', *American Economic Review*, Vol. 95, pp. 1–24.

Glaeser, E. L., Kominers, S. D., Luca, M. and Naik, N. (2018). 'Big data and big cities: the promises and limitations of improved measures of urban life', *Economic Inquiry*, Vol. 56, pp. 114–137.

Hanna, R. (2019). *New Research Busts the Myth of Welfare Dependency*, World Economic Forum, Geneva.

Hao, K. (2019). 'There's an easy way to make lending fairer for women. Trouble is, it's illegal', *MIT Technology Review*, Vol. 15, pp. 1–3.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd ed., Springer, New York.

HM Government. (2010). *State of the Nation Report: Poverty, Worklessness and Welfare Dependency in the UK*, HM Government, London.

Hoffman, M., Kahn, L. B. and Li, D. (2018). 'Discretion in hiring', *Quarterly Journal of Economics*, Vol. 133, pp. 765–800.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A. and Yarkoni, T. (2021). 'Integrating explanation and prediction in computational social science', *Nature*, Vol. 595, pp. 181–188.

Holm, E. A. (2019). 'In defense of the black box', *Science Magazine*, Vol. 364, pp. 26–27.

Hoynes, H. W. (1997). 'Work, welfare, and family structure: what have we learned?', in Auerbach A. J. (ed), *Fiscal Policy: Lessons from Economic Research*, MIT Press, Cambridge, pp. 101–146.

Hoynes, H. W. and Schanzenbach, D. W. (2012). 'Work incentives and the food stamp program', *Journal of Public Economics*, Vol. 96, p. 151.

Huang, L. Y., Hsiang, S. M., and Gonzalez-Navarro, M. (2021). *Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs*, NBER Working Paper 29105.

Ibrahim, R., Kim, S.-H. and Tong, J. (2021). 'Eliciting Human Judgment for Prediction Algorithms', *Management Science*, Vol. 67, pp. 1993–2656.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). 'Combining satellite imagery and machine learning to predict poverty', *Science Magazine*, Vol. 353, pp. 790–794.

Kang, J. S., Kuznetsova, P., Luca, M., and Choi, Y. (2013). *Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews*, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1443–1448.

Klapdor, M. (2013). 'Adequacy of income support payments', *Parliamentary Library Briefing Book*, Vol. 44, pp. 1–4.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015). 'Prediction policy problems', *American Economic Review: Papers & Proceedings*, Vol. 105, pp. 491–495.

Kleinberg, J., Lakkaraj, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017). 'Human decisions and machine predictions', *Quarterly Journal of Economics*, Vol. 133, pp. 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Rambachan, A. (2018). 'Algorithmic Fairness', *AEA Papers and Proceedings*, Vol. 108, pp. 22–27.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C. R. (2020). 'Algorithms as discrimination detectors', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 117, pp. 30096–30100.

Korpi, W. and Palme, J. (1998). 'The paradox of redistribution and strategies of equality: welfare state institutions, inequality, and poverty in the Western countries', *American Sociological Review*, Vol. 63, pp. 661–687.

Kretsedemas, P. (2005). 'Language barriers and perceptions of bias: ethnic differences in immigrant encounters with welfare system', *Journal of Sociology and Social Welfare*, Vol. 32, pp. 109–123.

Leana, C. R., Mittal, V. and Stiehl, E. (2012). 'Organizational behavior and the working poor', *Organization Science*, Vol. 23, pp. 888–906.

Lokshin, M. and Umapathi, N. (2022). *AI for social protection: Mind the people*, Brookings Institute, Washington, DC, pp. 1–5.

Madhusoodanan, J. (2021). 'A troubled calculus', *Science Magazine*, Vol. 373, pp. 380–383.

Markham, F. and Biddle, N. (2018). *Income*, Poverty and Inequality, Scientific Research Publishing, Canberra.

McBeath, B., Chuang, E., Bunger, A. and Blakeslee, J. (2014). ‘Under what conditions does caseworker-caregiver racial/ethnic similarity matter for housing service provision? An application of representative bureaucracy theory’, *Social Service Review*, Vol. 88, pp. 134–165.

McKenzie, D. and Sansone, D. (2019). ‘Predicting entrepreneurial success is hard: evidence from a business plan competition in Nigeria’, *Journal of Development Economics*, Vol. 141, 102369.

Mitrut, A. and Tudor, S. (2018). ‘Bridging the gap for Roma: the effects of an ethnically targeted program on prenatal care and child health’, *Journal of Public Economics*, Vol. 165, pp. 114–132.

Moffitt, R. (1985). ‘Unemployment insurance and the distribution of unemployment spells’, *Journal of Econometrics*, Vol. 28, pp. 85–101.

Moffitt, R. (1992). ‘Incentive effects of the US welfare system: a review’, *Journal of Economic Literature*, Vol. 30, pp. 1–61.

Mullainathan, S. and Spiess, J. (2017). ‘Machine learning: an applied econometric approach’, *Journal of Economic Perspectives*, Vol. 31, pp. 87–106.

Orlov, G., McKee, D., Foster, I. R., Bottan, D. and Thomas, S. R. (2021). ‘Identifying students at risk using a new math skills assessment’, *AEA Papers and Proceedings*, Vol. 111, pp. 97–101.

O’Sullivan, S., McGann, M. and Considine, M. (2019). ‘The category game and its impact on street-level bureaucrats and jobseekers: an Australian case study’, *Social Policy and Society*, Vol. 18, pp. 631–645.

Penman, R. (2006). ‘Psychosocial factors and intergenerational transmission of welfare dependency: a review of the literature’, *Australian Social Policy*, Vol. 2006, pp. 85–107.

Pentaraki, M. (2019). ‘Practising social work in a context of austerity: experiences of public sector social workers in Greece’, *European Journal of Social Work*, Vol. 22, pp. 376–387.

Price Waterhouse Coopers. (2016). *Baseline Valuation Report*, Department of Social Services, Canberra.

Price Waterhouse Coopers. (2019). *2018 Valuation Report*, Department of Social Services, Canberra.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z. and Mullainathan, S. (2019). ‘The algorithmic automation problem: prediction, triage, and human effort’, *ArXivWorking Paper*, Vol. 1903, pp. 1–20.

Raveendran, M., Puranam, P. and Warglien, M. (2022). ‘Division of labor through self-selection’, *Organization Science*, Vol. 33, pp. 810–833.

Reddel, T. (2018). *Using People-Centred Evidence to Shape Policy Strategy and Implementation*, Australian Institute of Family Studies, Victoria.

Reference Group on Welfare Reform. (2015). *A New System for Better Employment and Social Outcomes*, Reference Group on Welfare Reform, Canberra.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020). ‘Measuring the predictability of life outcomes with a scientific mass collaboration’, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 117, pp. 8398–8403.

Sansone, D. (2019). ‘Beyond early warning indicators: high school dropout and machine learning’, *Oxford Bulletin of Economics and Statistics*, Vol. 81, pp. 456–485.

Schmieder, J. F. and von Wachter, T. (2016). ‘The effects of unemployment insurance benefits: new evidence and interpretation’, *Annual Review of Economics*, Vol. 8, pp. 547–581.

Scoppetta, A. and Buckenleib, A. (2018). ‘Tackling long-term unemployment through risk profiling and outreach’, *European Commission – ESF Transnational Cooperation*, Vol. 6, pp. 1–28.

Staines, Z., Moore, C., Marston, G. and Humpage, L. (2021). ‘Big data and poverty governance under Australia and Aotearoa/New Zealand’s ‘‘social investment’’ policies’, *Australian Journal of Social Issues*, Vol. 56, pp. 157–172.

Stata (2019). ‘Cluster — Introduction to cluster-analysis commands’, in *Stata Multivariate Statistics Reference Manual* (Release 16), Stata Press, College Station, pp. 100–111.

Stevenson, M. T. and Doleac, J. L. (2019). *Algorithmic Risk Assessment in the Hands of Humans*, IZA Discussion Paper Series No. 12853, 1–74.

Tseng, Y.-P. and Wilkins, R. (2003). ‘Reliance on income support in Australia: prevalence and persistence’, *Economic Record*, Vol. 79, pp. 196–217.

UnitingCare Australia. (2021). *Submission to the Senate Standing Committees on Community Affairs into the Social Services Legislation Amendment (Strengthening Income Support) Bill 2021*.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E. and Jiang, N. (2013). 'Children in the public benefit system at risk of maltreatment: Identification via predictive modeling', *American Journal of Preventive Medicine*, Vol. 45, pp. 354–359.

Van Landeghem, B., Desiere, S. and Struyven, L. (2021). 'Statistical profiling of unemployed jobseekers', *IZA World of Labor*, Vol. 483, pp. 1–9.

Vooren, M., Haelermans, C., Groot, W. and van den Brink, H. M. (2019). 'The effectiveness of active labor market policies: A meta-analysis', *Journal of Economic Surveys*, Vol. 33, pp. 125–149.

Welfare Working Group. (2011). *Reducing Long-Term Benefit Dependency: Recommendations*, Welfare Working Group, Auckland.

Whiteford, P. (2010). 'The Australian Tax-Transfer System: Architecture and Outcomes', *Economic Record*, Vol. 86, pp. 528–544.

Whiteford, P. (2018). 'Why social policy counts', *Inside Story*, Vol. November, pp. 1–7.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S. and Burke, M. (2020). 'Using publicly available satellite imagery and deep learning to understand economic well-being in Africa', *Nature Communications*, Vol. 11, p. 2583.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Data S1** Supporting Information.