Predicting Disability Insurance Entry and Long-Term Dependence: A Machine Learning Approach

Guida Ayza Estopà* (DULBEA, ULB) & Ilan Tojerow (DULBEA, ULB)
*guida.ayza.estopa@ulb.be
September 2025

Disability Insurance (DI) recipiency has risen sharply in Belgium. Using administrative data and machine learning techniques, we predict both DI entry and the transition to long-term dependence. Observable characteristics explain about one third of entry risk and one quarter of persistence, with health variables and prior DI spells driving entry, and age and mental health shaping long-term outcomes. A small group of individuals concentrates most of the predicted risk, highlighting the potential of predictive models to inform preventive policy.

JEL classification: H53, H55, I38, J14, C55
Key words: Welfare policies, Disability Insurance, Prediction, Machine Learning.

1. Introduction

Disability Insurance (DI) constitutes a cornerstone of social protection in developed economies, standing among the largest and most significant welfare programs. Over the past decades, the number of DI beneficiaries has increased substantially across OECD countries, raising concerns about the underlying drivers of this growth. While some countries have managed to stabilize or even reduce participation rates, Belgium has experienced a striking rise: the share of the working-age population receiving DI benefits for more than one year almost doubled, from 3.5% in 2005 to 6.8% in 2020 (INAMI, 2022). Figure 1, Panel A, documents the sharp growth in the stock of long-term DI recipients between 2009 and 2020, whereas Panel B highlights the surge in annual inflows into short-term DI, particularly since 2015. This expansion has turned Belgium into one of the OECD countries with the highest levels of public expenditure on incapacity. As shown in Figure 2, Belgium records the second highest level of cash spending on DI, and when combining cash with in-kind benefits, its spending ranks just behind Nordic countries such as Norway, Denmark, and Sweden. This corresponded to 3.5% of GDP in 2020 (OECD, 2025).

While DI is crucial for protecting individuals facing long-term health limitations, the rapid expansion of beneficiaries also raises concerns about sustainability. Prolonged reliance on DI can strain public finances and contribute to the social risks associated with labour market detachment. In light of these challenges, policy discussions and research have increasingly studied preventive strategies, aiming to reduce the incidence of new DI entries in the first place, while understanding its determinants. Several papers have analysed specific determinants of DI entry and long-term dependence—such as the effects of financial incentives through taxes or benefit generosity (Maestas et al., 2013; Kostol & Mogstad, 2021; Marie & Vall-Castelló, 2023), policy interventions applied to the DI or to other welfare programs (Campolieti & Riddel, 2012; Fontenay & Tojerow, 2025; De Brouwer et al., 2023), and, to a lesser extent, other dimensions such as gender differences (Low & Pistaferri, 2019), administrative timings (Autor et al., 2015), and information frictions (Kostol & Myhre, 2021). However, there is limited research providing a comprehensive perspective that simultaneously integrates labour, health, and socioeconomic factors. The literature has emphasized a broad set of mechanisms shaping DI recipiency rates. In the U.S., Autor and Duggan (2003, 2006) document how institutional reforms and changing labour market incentives drove much of the expansion of DI, while Duggan and Imberman (2009) show that population aging has played only a minor role. Similar evidence emerges across other OECD countries, where Börsch-Supan et al. (2009) find that health indicators explain little of the cross-country variation, whereas eligibility rules and replacement rates are crucial. More generally, there is broad consensus that demographic and health factors cannot account for more than half of the observed growth in DI recipiency (Burkhauser et al., 2014; Liebman, 2015).

Labor market conditions also play a central role. Charles et al. (2018) highlight how weak labour demand increases DI take-up, while Maestas et al. (2013) and Carey et al. (2022) show that inflows rise during recessions, making it difficult to disentangle the relative contribution of health shocks and cyclical downturns. These dynamics underscore the sensitivity of DI to both structural and temporary changes in employment opportunities. More recently, attention has shifted toward job quality and psychosocial risks as emerging drivers of DI entry. Rising cases of burnout illustrate this trend—DI claims due to burnout in Belgium increased by 40% between 2015 and 2021 (INAMI, 2023). Empirical evidence confirms that poor working conditions and work-related stress are strong predictors of DI entry (Clumeck et al., 2009; Holmgren et al., 2012; Nekoei et al., 2025).

When focus on the Belgian case, Saks (2017) suggests that much of the recent increase in recipiency can be attributed to higher labour force participation among women and older workers. However, De Brouwer and Tojerow (2023) provide a more comprehensive analysis that accounts for multiple dimensions—demographics, labour market participation, and job characteristics—and show that changes in observable characteristics, such as gender, age structure and job characteristics, can only marginally explain the increase in long-term DI entry.

The present paper leverages machine learning (ML) techniques to identify how different sets of observed factors—labour market histories, health-related determinants, and socioeconomic characteristics— and unobserved factors, interact in shaping DI trajectories, thus identifying the most relevant predictors of DI risk. By using ML, we can capture these complex relationships, improve out-of-sample prediction, and offer a more flexible and data-driven approach to understand DI dynamics. Our analysis builds on the framework of Mueller & Spinnewijn (2023), which examines unemployment insurance (UI) dynamics through the lenses of heterogeneity, dynamic selection, and duration dependence. We extend these concepts to the case of DI, exploring how pre-existing individual differences, changes in the risk pool over time, and time trends influence both DI entry and the transition from short-term to long-term DI. Specifically, we seek to answer three key questions: First, how much heterogeneity on the probability to enter DI is observed? Second, what are the primary drivers of these observed probabilities? Do health-related variables, such as mental health drug prescriptions or frequency of doctor visits, play a dominant role, or are labour market histories and socioeconomic characteristics equally predictive? And third, is there persistence in the predictive power through the years? Or may business cycles influence the composition of individuals to enter DI? Moreover, we are going to answer the same questions for a second outcome explaining the determinants of whether an individual on short-term DI transitions to long-term DI. Is there any observable heterogeneity in the individual characteristics that increases the risk of, not only enter on DI, but also being long-term disabled? Understanding these dynamics is crucial not only for policymakers aiming to design more efficient social insurance systems but also for improving predictive models of DI risk.

Our analysis relies on a rich dataset that combines administrative records on labour market history, healthcare utilization, and sociodemographic characteristics from two main sources: the Data Warehouse from the Social Security (DWSS-BCSS) and the Inter-mutualist Agency (IMA-AIM) in Belgium. The dataset spans the period from 2006 to 2019, covering the 10% of the Belgian population, which consist of 735,000 individuals with quarterly observations totalling almost 40 million observations. The paper starts with a conceptual framework illustrating how different sources of heterogeneity, both observed and unobserved, may influence DI entry rates as well as the probability of remaining on DI for a long period. DI entry and long-term dependence are shaped by dynamics similar to those in UI, particularly through heterogeneity in initial risk, dynamic selection pre-DI, and duration dependence. Heterogeneity in DI risk arises from individual differences in observable characteristics such as age, gender, income, and health status (proxied through various dimensions of health expenditure), as well as unobservable traits like resilience or health-seeking behavior. Over time, dynamic selection occurs as individuals with higher health risks are more likely to enter DI, changing the composition of the remaining population not on DI. This process is comparable to the dynamics in UI, where those with higher employability exit the system earlier. Duration dependence refers to the likelihood of transitioning from short-term to long-term DI, which may increase for some individuals with certain observed or unobserved characteristics. While a rich literature has already documented substantial heterogeneity in job finding rates among unemployed workers (Álvarez & Shimer, 2011; Cockx et al., 2023), there is little evidence on the extent of such heterogeneity in DI outcomes. These factors point out how machine learning models, which can capture complex, nonlinear interactions, offer significant advantages over traditional econometric approaches in predicting DI outcomes.

For the empirical analysis we employ standard Machine Learning (ML) techniques, training a prediction model on a training sample and then evaluating the predictive power in a hold-out sample. The main problem in all prediction exercises is the trade-off between improving the prediction model and overfitting it when including too many variables. ML methods and the separation of the two samples help to optimize variable selection and to deal with the over-fitting problem in a data-rich environment. We focus on two outcomes: (i) the probability of entering DI, and (ii) the likelihood of transitioning from short-term to long-term DI. We define these probabilities as the risk variables for our model. We use two ML models: Random Forest and Gradient Boosted Regression Trees and combine them in an Ensemble Model, which is a linear weighted combination of them. After tunning the models in the training sample, we estimate them to obtain the ensemble predictors and the calibrated probabilities for each outcome. We are going to apply this year by year to analyse time differences.

Our results show that machine learning models can capture a substantial share of the heterogeneity in Disability Insurance outcomes. For DI entry, the baseline model achieves strong predictive accuracy (AUC ≈ 0.91), explaining about one third of the variation in entry risks. Health-related variables emerge as the most informative predictors, followed by labour

market histories, while sociodemographics alone contribute very little. A particularly strong factor is prior DI spells, highlighting the recurrence of DI dependence, although predictive power remains high even when restricting the analysis to first-time entrants. For the transition from short-term to long-term DI, the model explains close to one quarter of the variation, with age, psychiatric consultations, and pharmaceutical expenditure ranking among the top predictors, underscoring the distinct role of mental health and weaker labour market attachment in persistence. The distribution of predicted probabilities further illustrates how a small group of individuals concentrates most of the risk, especially once already on DI. Ongoing work extends these analyses by examining robustness across subsamples, the evolution of predictive power over elapsed time and forecasting horizons, and the stability of results across years and business-cycle conditions.

This paper contributes to the literature in several ways. First, it is the first to systematically document heterogeneity in DI risks, both at entry and in the transition from short- to long-term dependence, an aspect that has been largely overlooked in previous research. Second, by applying machine learning methods to rich administrative data, we are able to capture complex, nonlinear interactions and quantify how much of the observed heterogeneity can be explained by observable factors. Third, we provide new evidence on the dynamic dimension of DI risk, showing how predictability evolves with spell duration, forecasting horizons, and business-cycle conditions. Taken together, these contributions open new perspectives for preventive policy design by identifying the profiles most at risk of DI dependence.

The paper is organized as follows. Section 2 describes the data and the institutional context. Section 3 provides a conceptual framework of the heterogeneity in DI risk, the dynamics of the probabilities of being on DI, and duration dependence. Section 4 explains the methodology used. Section 5 present the findings. Section 6 concludes.

2. Data and Context

The present analysis relies on a rich administrative dataset combining longitudinal records on labor market history, healthcare utilization, and sociodemographic characteristics from two main sources: the Data Warehouse of the Social Security (BCSS-DWSS) and the Inter-Mutualist Agency (IMA-AIM) in Belgium. The data span the period from 2006 to 2019 and cover 10% of the Belgian working-age population, which consist of 735,000 individuals with quarterly observations, totaling around 40 million records. Data from different sources are linked using an anonymized social security identifier.

We use several datasets from the BCSS, which is a central data system managed by the Belgian government that integrates individual-level administrative records across social security,

healthcare, and employment domains. Our main variables of interest are those related to DI. In this context, we have data on short- and long-term disability spells, as well as related information such as the type of benefits, household composition, and, in the case of long-term DI, the pathology that led to the disability. In Belgium, workers who have contributed sufficiently to the social insurance system are eligible for DI if they are unable to work for health-related reasons, regardless of their employment status at the onset (employed or unemployed). During the first month of sickness, white-collar workers receive full salary financed by the employer, whereas for blue-collar workers, the employer covers part of the salary, and the rest is paid by the National Institute for Health and Disability Insurance (NIHDI). From the second month onwards, all benefits are paid by the individual's health insurance fund (mutuality). The replacement rate depends on prior employment status and declines with the duration of the sickness spell, starting at 60% of gross salary for most workers, subject to income-dependent caps and floors.

Sick leave can be prescribed by any treating medical practitioner from the first day of illness. To qualify, three conditions must be met: (i) the individual must cease all productive activity; (ii) the cessation must be due to a deterioration in health unrelated to professional activity; and (iii) work capacity must be reduced by at least 66% relative to their previous occupation. After one month, an advisory physician evaluates eligibility for short-term DI. After one year, the mutuality doctor may propose a transition to the long-term DI scheme, which requires a reassessment by a certified medical advisor. Long-term DI benefits are also financed by the NIHDI and differ mainly in how residual work capacity is assessed and in the replacement rate, which increases to 65%, adjusted for household composition. We define two main outcome variables: starting a DI spell and transitioning from Short-term DI to LT-DI. We identify an individual as starting a DI spell when they start receiving benefits from their mutuality, so just after the first month which is covered by the employer, and transition to a LT-DI when they start receiving the other type of benefits because they already spent one year disabled.

As predictor variables, we first include those related to the labour market, which are also drawn from the BCSS. These include worker type (blue- or white-collar; public or private sector), working time (part- or full-time), self-employment status, unemployment spells, and income data categorized into normalized wage brackets. In Belgium, salaried workers are eligible for unemployment insurance (UI) following involuntary job loss, conditional on a sufficient employment history. Uniquely, UI benefits can be received for an indefinite period, provided the claimant remains available for the labour market and complies with job search obligations and reintegration plans. The benefit amount decreases progressively over time, depending on past earnings, family situation, and unemployment duration. A minimum benefit is guaranteed, but non-compliance may lead to benefit suspension or reduction. Self-employed individuals are subject to different eligibility rules for both DI and UI. The BCSS

data also include sociodemographic information such as age, gender, nationality, household composition, and region of residence, that are also included as predictor variables.

The second main data source is the IMA-AIM, which collects and harmonizes individual-level healthcare and reimbursement data from all Belgian mutual insurance funds. In Belgium, all residents must be affiliated with a mutuality, which acts as an intermediary between individuals and the compulsory public health insurance system. Mutualities, apart from the payment of health-related benefits, also manage the reimbursement of medical expenses. Although they are private entities with voluntary affiliation, they operate under a public and regulated framework.

From the IMA, we obtained data on reimbursements—hence consumption—of prescription drugs (whether purchased in public pharmacies or administered in hospitals) and other health expenditures such as general practitioner and specialist visits, as well as hospital stays. Drug information is categorized using the ATC (Anatomical Therapeutic Chemical) classification. At its first level, the ATC identifies the anatomical system targeted by the drug. We focus in particular on drugs affecting the nervous and musculoskeletal systems, which include medications for mental health and musculoskeletal disorders respectively, the two main conditions leading to DI spells. For these two categories, we have more granular data up to ATC level 3, which allows us to distinguish, for instance, between antidepressants and antipsychotics within the nervous system category.

Table 1 presents the full set of variables used in the baseline prediction model, which are generally available for all quarters in the sample. All predictor variables are measured prior to the predicted event—that is, before the onset of the DI spell or the observed transition to LT-DI spell. For historical variables, we use a two-year lookback window from the year of analysis (e.g., the 2018 unemployment history variable indicates whether the individual was unemployed at any point since 2016).

The model includes three main groups of predictors. First, sociodemographic variables: gender, age, marital status, number of children, nationality of origin, region, and, where available, district of residence. Second, health-related variables, such as whether the individual had any visits to a general practitioner or specialist in the previous two years, as well as the number of such visits. Given the prevalence of mental health and musculoskeletal disorders among DI recipients, we also include visits to psychologists, psychiatrists, and physiotherapists. In addition, we consider the number of hospitalization spells and total days spent in hospital, allowing us to distinguish between short recurrent hospitalizations and prolonged stays. Third, we incorporate pharmaceutical variables, focusing on drug consumption related to mental health and musculoskeletal conditions. Specifically, we include indicators for any drug belonging to the ATC level 1 categories corresponding to the nervous system and musculoskeletal system, respectively. In addition, we separately identify the use

of antidepressants (ATC level 3). For each of these drug groups, we distinguish whether the drug was purchased in a public pharmacy or administered in a hospital. For all variables, we record both whether the drug was taken at least once and the total quantity consumed. Finally, we include labor market variables to assess the role of employability on the probability of being on DI. These include unemployment or self-employment status in the two years prior to the disability spell, working time (full- or part-time), type of occupation (blue- or white-collar), employment sector (public or private), and labor income. Income is reported in normalized brackets and refers to total earnings in the previous completed calendar year; while less precise than raw earnings, it allows for meaningful comparisons across individuals.

We also include an indicator for whether the individual had any prior disability spells. This variable is used both as a predictor and, in an alternative specification, to restrict the sample to individuals with no previous DI episodes, in order to analyse the determinants of first-time entries into DI.

Table 2 presents descriptive statistics for the overall working-age population and for the subpopulation of individuals currently on DI. In the full sample, the gender, nationality, and age distributions are balanced, with 50.1% women, 22.2% foreign nationals, and a mean age of 40.9 years. Regarding mental health–related healthcare use, 34.4% of individuals have at some point purchased medication in a public pharmacy and 7.5% have received such medication in a hospital. General practitioner visits are nearly universal (95.1%), and more than half of the sample (51.6%) has experienced at least one hospital stay. In terms of labour market histories, 6.8% of individuals had an unemployment spell, 6.5% ever entered DI, and 2.7% transitioned to long-term disability.

The DI subpopulation displays broadly similar demographic characteristics, though with slightly higher mean age (44.8 years) and a lower share of women (40.5%). Differences are more pronounced in healthcare use: 49.9% have purchased medication in a public pharmacy and 13.7% have received it in a hospital, with almost all individuals having visited a general practitioner—reflecting their central role in issuing DI certificates—and a large majority having experienced a hospital stay (78.0%). Labour market differences are also notable: unemployment spells are nearly twice as frequent among individuals on DI (11.4% compared to 6.8% in the overall population). 7.4% of this group transitioned to long-term disability.

3. Conceptual framework

We aim to explain why the number of people on DI has increased so much on the last decade. With the models used in this paper, we are going to study which family of factors is better at predicting the entrance into DI. In a second exercise we are going also to understand the

predicting factors for being long term disabled (to transition from short term DI to long term DI). Identifying these micro-level determinants may help us understand the mechanisms behind the aggregate trend.

Following Mueller & Spinnewijn, we present a conceptual framework to account for heterogeneity in DI risk, the dynamics of the probabilities of being on DI, and duration dependence. Their work builds on the unemployment benchmark, where job-finding dynamics are well established. In our case, we need to build the link between individual characteristics—socioeconomic, labor-related, and health-related—and the probability of transitioning into DI.

We first describe heterogeneity in the initial DI risk and then turn to dynamic selection, duration dependence, and other factors that may influence DI hazards.

3.1. Heterogeneity on the initial DI risk:

The first step is to define how different sources of heterogeneity, both observed and unobserved, may influence DI entry rates as well as the probability of remaining in DI for a long period. De Brouwer & Tojerow (2023) offer an extensive analysis of DI determinants in Belgium, showing that changes in observable characteristics such as age and work type only marginally account for the increase in the long term. Heterogeneity in DI risk arises from differences in observable characteristics that affect baseline DI risk—such as age, gender, income, labor market position, and health status (which we proxy through dimensions of healthcare use)—as well as unobservable traits such as resilience, health-seeking behavior, and moral hazard.

It is widely accepted that age, gender, and income play an important role in the probability of starting a DI spell (INAMI, 2028; De Brouwer & Tojerow, 2023). Health is by definition negatively correlated with DI entry, since eligibility requires reduced work capacity due to a health condition. At the same time, certain work characteristics can make continued employment feasible despite health limitations, while others may increase the likelihood of health deterioration and DI entry—for example, stress or burnout (Moreau et al., 2004; Toppinen-Tanner et al., 2005; Holmgren et al., 2013; INAMI 2023, INAMI-AIM, 2024).

The literature has focused strongly on ageing and the increase in women's employment, but we consider a broader set of observables grouped into three domains: sociodemographic factors, labor-related factors, and health-related factors. Whereas health is often proxied simply by age, we use healthcare consumption to provide a more direct measure of health status.

However, we are also aware of the existence of unobservable heterogeneity; individual resilience, health-seeking behavior, moral hazard, and work preferences affect DI risk beyond what can be explained by observables. Moral hazard here refers to the behavioral response to the incentives created by the DI system, such as a greater tendency to apply for benefits or a lower probability of returning to work when benefits reduce the financial cost of non-employment (Autor & Dugan, 2003; Maestas et al., 2013; French & Song, 2014; Kostøl & Mogstad, 2014).

Formalization:
Formally, we write the individual DI entry probability as

$$D_{i,t} = D_t(X_i) + \varepsilon_{i,t}$$

where $D_t(X_i) = E_t(D_{i,t}|X_i)$ is the individual DI entry probability based on observable characteristics X_i in time t, and $\varepsilon_{i,t}$ captures unobservable characteristics, assumed orthogonal to the observables:

$$E_t(\varepsilon_{i,t}|X_i) = 0$$

The set of observable characteristics is:

$$X_i = \{S_i, L_i, H_i\}$$

where S_i refers to sociodemographic variables, L_i labour market characteristics, and H_i health-related factors.

We cannot observe individual DI entry probabilities directly, but we do observe whether an entry occurs. The realization of the probability is:

$$R_{i,t} = \begin{cases} 1 & with\ probability\ D_{i,t}, \\ 0 & otherwise, \end{cases} \qquad D_{i,t} \in (0,1)$$

Thus, our prediction model estimates the probability of $R_{i,t} = 1$ based on X_i:

$$R_{i,t} = R_t(X_i) + e_{i,t}$$

where $e_{i,t}$ is the prediction error. If the prediction model is unbiased, then

$$E_t(R_{i,t}|X_i) = D_t(X_i)$$

Proposition 1. Lower bound on heterogeneity
We are first interested in quantifying the extent of observable heterogeneity in DI entry. In practice, the explanatory power of the prediction model is summarized by the R^2, which

measures the share of the variance in realizations $R_{i,t}$ that can be explained by observables $X_i = \{S_i, L_i, H_i\}$.

Formally, for a hold-out sample,

$$R^2 = 1 - \frac{\sum_i (R_{i,t} - \hat{R}_{i,t})^2}{\sum_i (R_{i,t} - \bar{R}_{i,t})^2}$$

where $\hat{R}_{i,t} = R_t(X_i)$ is the predicted probability of entry and $\bar{R}_{i,t}$ the sample mean.

This measure represents a lower bound on heterogeneity because it captures only the variation explained by the included observables. Unobservable factors $\varepsilon_{i,t}$ remain outside the model, so the R^2 cannot capture the full extent of heterogeneity in DI entry.

By comparing R^2 across different sets of variables (e.g. only sociodemographics, then adding labour, then adding health-related variables), we can document how each block contributes to explaining DI risk. In particular, healthcare consumption is expected to be especially informative, as it proxies underlying health status much more directly than age or other demographic controls.

Proposition 2. Persistent heterogeneity

We next ask whether heterogeneity is persistent (time-invariant) or transitory (time-varying). Persistence would indicate that certain risk factors (e.g. gender, education, permanent health conditions) produce stable differences in DI risk across years, while transitory components would capture factors that change over time, such as health shocks, business-cycle conditions, or short-term institutional effects.

Formally, let the variance of the individual DI risk at time *t* be:

$$Var_t(D_{i,t}) = Cov_t(D_{i,t}, D_{i,t'}) + Cov_t(D_{i,t}, D_{i,t} - D_{i,t'}),$$

where t' refers to another period.

- The first term on the right side of the equation captures the persistent component, i.e. the part of the variance that remains stable across periods.
- The second term captures the transitory component, i.e. the part of the variance that changes between periods.

A high covariance indicates that the same individuals are consistently predicted at high (or low) risk across periods, i.e. persistent heterogeneity. A low covariance would suggest that heterogeneity is mainly transitory.

Equivalently, we can compute the cross-year R^2: the explanatory power of predictions from year t' when applied to realizations in year t. This corresponds to:

$$R^2_{t,t'} = Corr(D_t(X_i), D_{t'}(X_i))^2$$

which measures how stable predicted risks are over time.
In practice, this involves estimating prediction models year by year and then examining how much predictive power transfers across years.

3.2. Selection on the probability to be on Long-Term DI

We next focus on the second outcome of interest: the probability that individuals in short-term DI transition to long-term DI after twelve months. Formally, for the set S_t of individuals in ST-DI at year t, we define:

$$E_{i,t+1} = \mathbf{1}\{\text{transition to } LTDI \text{ at } t+1\}, \quad i \in S_t$$

Where $\mathbf{1}\{\cdot\}$ denotes the indicator function, equal to 1 if the condition is satisfied and 0 otherwise, and estimate the transition probability conditional on observables at t:

$$D_{i,t+1} = D_{t+1}(X_{i,t}) + \varepsilon_{i,t+1}, \qquad \mathbb{E}[\varepsilon_{i,t+1}|X_{i,t}] = 0$$

Our prediction model provides $\hat{p}_{i,t+1} = \hat{D}_{t+1}(X_{i,t})$, and predictive performance in a hold-out sample is summarized by

$$R^2_{t\to t+1} = 1 - \frac{\sum_{i\in S_i}(E_{i,t+1} - \hat{p}_{i,t+1})^2}{\sum_{i\in S_i}(E_{i,t+1} - \bar{E}_{t+1})^2}$$

This captures how much of the heterogeneity in the ST-DI to LT-DI transition can be explained by observables. A higher R^2 indicates that the pool of individuals reaching the 12-month threshold is increasingly composed of systematically high-risk individuals (individuals with poorer health, weaker labor market attachment, or stronger attachment to the DI scheme), consistent with dynamic selection.

3.3. Duration dependence vs. dynamic selection

Finally, we distinguish between true duration dependence and dynamic selection in the evolution of DI spells. Duration dependence refers to changes in an individual's hazard of exit or transition as the spell lengthens, while dynamic selection arises because, over time, those who remain in DI are disproportionately individuals with lower exit probabilities.

Formally, let $h_i(d)$ be the exit probability of individual i from ST-DI at elapsed duration d. The observed hazard at duration d is the average across those still at risk,

$$\bar{h}(d) = \mathbb{E}[h_i(d)|i \in S_d]$$

The change in the observed hazard between d and d+1 can be decomposed into:

$$\bar{h}(d+1) - \bar{h}(d) = \mathbb{E}[\bar{h}_i(d+1) - \bar{h}(d) \ |i \in S_{d+1}] + (\ \mathbb{E}[h_i(d)|S_{d+1}] - \mathbb{E}[h_i(d)|S_{d+1}])$$

The first term is the duration dependence, and it captures genuine changes in individual hazards as the spell progresses—for instance, because health deteriorates, labor market attachment weakens, or adaptation to DI increases. The second term reflects the dynamic selection: individuals with higher exit probabilities tend to leave earlier, so those who remain are increasingly concentrated among the hardest-to-exit cases.

In practice, we will assess duration dependence by studying how predictability evolves as the spell advances. In addition, we will evaluate predictive performance at alternative horizons (2- and 5-year risks), which complements the duration analysis by providing a medium-run perspective on DI risks. Finally, we will also examine how predictability varies across calendar years. Two empirical extensions are therefore planned: (i) comparing predictive performance at different elapsed quarters within ST-DI, and (ii) repeating the analysis across years to capture differences linked to business-cycle conditions.

4. Methodology

For the empirical analysis we employ standard Machine Learning (ML) techniques, training a prediction model on a training sample and then evaluating the predictive power in a hold-out sample. The main problem in all prediction exercises is the trade-off between improving the prediction model and overfitting it when including too many variables. ML methods and the separation of the two samples helps to optimize variable selection and to deal with the over-fitting problem in a data-rich environment. We focus on two outcomes: (1) the probability of entering DI, and (2) the likelihood of transitioning from short-term to long-term DI. We define these probabilities as the risk variables for our model. We start using three ML models: Random Forest, Gradient Boosted Regression Trees and LASSO in the baseline model for the year 2018, after analyzing their discriminatory power and overall performance in the training sample, we decide to keep only the two first models and combine them in an Ensemble Model, which is a linear weighted combination of them. These models take different approaches for the selection of variables, but also allow differently for nonlinearities and interactions between these variables. Random Forest tends to be more robust to noise and provides stable predictions across many weakly correlated trees, while Gradient Boosted Trees sequentially focus on harder-to-predict cases and typically achieve higher accuracy by minimizing residual errors. Combining both leverages the strengths of each method.

We divide the sample in a 60% training sample and a 40% hold-out sample. The first step of the prediction process, after preparing the data and selecting the variables, is tunning key parameters for all the prediction models, we follow standard practice in machine learning and do it by 3-fold cross-validation. We then estimate the different models separately, obtain the Ensemble Model and calibrate the probabilities for each outcome. Later on, we apply this year by year to analyze time differences.

For the tunning process, we use the 15% of the sample to optimize, among other features, the minimal node size and the number of variables used at each node for the Random Forest model, the learning rate for the Boosted Regression Trees, and the shrinkage parameter for the LASSO. We run and compare different alternatives, and we finally choose the one that optimizes the area under the receiver operating characteristic curve (ROC-AUC), which is also a standard practice in ML. Within this process, we run the models several times adapting the parameters to the ones that result in a better performance, this includes changing the hyperparameters of the algorithm, the sampling technique, the validation process and the set of predictor variables. It also involves a deep study of the contribution of the variables and their possible interactions between them. Once decided the best tunning parameters, the three models are estimated using a 30% of the sample not used before. As a third step in the prediction model, we use 7.5% of the sample to obtain the Ensemble Model. Instead of a simple weighted combination, we apply a stacking approach, where a logistic regression model learns to optimally combine the predictions from the random forest and the gradient boosting regression trees. The probability we get from the Ensemble Model can be defined as:

$$p_{EnsembleModel} = \beta_{RF}\hat{p}_{RF} + \beta_{GB}\hat{p}_{GB}$$

where $\hat{p}_x$ is the prediction from algorithm x and β_x is the associated weight. Finally, we calibrate the raw predictions get from the ensemble model to the actual observed probabilities by estimating a linear spline in a different 7.5% of the sample. This flexible functional form allows for piecewise linear adjustments to better align predicted and observed risks. After these steps, we evaluate the final model on a hold-out sample, which represents 60% of the data. This sample has not been used in any previous step, ensuring an unbiased assessment of the model's performance. The results we present correspond to this hold-out evaluation for the year 2018. We provide more details on the Ensemble Model, the tuning process and the estimation of the underlying prediction algorithms in the Appendix.

To evaluate the accuracy of our prediction model, we compare predictions and outcomes in the hold-out sample for the year 2018 in Figure 3. Panel A displays results for outcome 1, the probability of entering DI. Individuals are grouped into 10 equally sized bins based on predicted risk. For each bin, we plot the average predicted probability against the observed DI entry rate. The dashed 45-degree line indicates perfect calibration. The points lie close to this line, indicating that the model's predictions are well aligned with actual outcomes. Panel B

shows analogous results for outcome 2, the probability of transitioning from short-term to long-term DI. As with outcome 1, predictions track observed rates reasonably well. In both panels, predicted probabilities remain below 0.4 across all bins. This upper bound reflects the underlying incidence of DI events, which is relatively low in the population: only a limited share of workers ever enter DI, and an even smaller fraction transition from short- to long-term DI. Consequently, even those at highest predicted risk face probabilities well below one. Far from indicating poor performance, this pattern highlights the model's ability to capture meaningful variation in risk within the empirically relevant range. The close alignment between predicted and observed rates suggests that the model is well calibrated and provides reliable risk stratification despite the inherently low baseline probabilities. In both panels, predicted probabilities remain below 0.4 across all bins. This upper bound reflects the relatively low incidence of DI events: only a limited share of workers enter DI, and an even smaller fraction transition from short- to long-term DI. Consequently, even individuals at highest predicted risk face probabilities well below one. Within this empirically relevant range, the model captures substantial variation in risk and shows close alignment between predicted and observed rates, indicating good calibration and reliable predictive performance. In Appendix, we further assess calibration by performing the evaluation separately for different subgroups. Specifically, we split the sample by gender, age, pathology, and salary levels (work in progress), in order to examine whether predictive performance is consistent across heterogeneous groups and to rule out systematic biases in specific segments of the population.

Assessing the model:

We use our prediction model to assess the probability of entering DI and, conditional on entry, the probability of transitioning to long-term DI. Figure 4 displays the distribution of calibrated predicted probabilities in the 2018 hold-out sample. Panel A shows the distribution for DI entry among the full population. The probabilities are highly concentrated near zero, reflecting the fact that only a small share of individuals actually enters DI. This concentration illustrates the challenge of predicting a relatively rare event. Panel B shows the corresponding distribution for transitions to long-term DI among those who already entered DI. Again, the mass of the distribution lies at the lower end, consistent with the relatively low incidence of long-term transitions.

Despite the rarity of both outcomes, the model performs well when evaluated on predictive accuracy. Figure 5 reports the Receiver Operating Characteristic (ROC) curves. The ROC curve contrasts the true-positive rate with the false-positive rate at different thresholds for classifying predicted probabilities into binary outcomes. The area under the curve (AUC) equals 0.907 for DI entry and 0.72 for long-term transitions, compared to a benchmark of 0.5 for random guessing and 1 for perfect prediction. These values indicate excellent discriminatory power for DI entry and solid performance for transitions to long-term DI.

As an additional measure, we compute the R-squared, which equals 0.32 for DI entry and 0.073 for long-term transitions. While these values may seem low, this is expected in models with binary outcomes, where the dependent variable is a random realization of an underlying probability. In this context, the reported values still indicate that the model captures a substantial share of the systematic variation in DI risks.

Predicting variables:
We are interested in how different sets of variables contribute to the predictive power of our model. Prior work has highlighted the role of observable characteristics such as age and gender in shaping DI risks. De Brouwer and Tojerow (2023) show that demographic shifts and labour market participation contributed to the rise in DI recipiency in Belgium between 2005 and 2020, but that these factors alone account for only a limited part of the overall increase. This suggests that additional sources of heterogeneity are important for explaining DI entry.
The strength of our analysis lies in the data-rich environment, which allows us to incorporate a wide set of sociodemographic, labour market, and health-related characteristics beyond those considered in previous studies. Table 1 summarizes the groups of variables included in our models. To evaluate their relative contribution, we compare a series of nested models. Starting from a baseline specification that only includes sociodemographic information, we sequentially add labour market characteristics, healthcare consumption, and finally the full set of predictors.

To further formalize the contribution of different sets of variables, we add them separately to the sociodemographic baseline. Table 4 presents these results (ongoing). Although the ordering of variables depends on correlations between predictors, the analysis confirms that health-related variables are the key drivers of predictive performance, while labour market and sociodemographic factors contribute more modestly.

5. Findings

5.1. Risk of entering DI

Our baseline model achieves an AUC of 0.91, confirming that the predictions discriminate well between individuals who do and do not enter DI. More importantly, the model attains an R^2 of 31.6%, meaning that roughly one third of the heterogeneity in DI entry is explained by observable characteristics. This is a substantial share in the context of a rare and multifactorial outcome such as DI. The remaining variation reflects unobservable dimensions, such as health-seeking behaviour, individual resilience, preferences for work, or behavioural responses like moral hazard. Rich administrative data therefore provide meaningful insights into the distribution of DI risk, while unobserved factors continue to play a complementary role.

Looking at the distribution of predicted probabilities gives further perspective. Most individuals are concentrated near zero risk, consistent with the low incidence of DI entry, while a small minority is clearly separated with higher predicted risks (Figure 4, Panel A). This unequal distribution shows that observable characteristics allow us to identify a limited group of workers who concentrate most of the predicted risk, while the majority remain at negligible risk.

To better understand which characteristics, account for the observed heterogeneity in DI entry, we compare models that sequentially add sociodemographic, labour market, and health-related variables. The results in Table 3 show that sociodemographics on their own provide very limited explanatory power (R^2=0.025, AUC = 0.65). Adding labour market information substantially increases predictability (R^2=0.088, AUC = 0.79), while health variables are even more informative, raising the fit to R^2=0.21 and the AUC to 0.86. This decomposition highlights that medical information, proxied by healthcare consumption, contributes most to explaining entry into DI, whereas demographics by themselves account for only a small share of the variation. "Labour market factors provide additional predictive content, but their contribution remains smaller than that of health-related variables.

Figure 6, Panel A, provides further detail by reporting variable-importance scores for the full model. The dominant predictor is whether the individual has experienced a previous DI spell, underscoring the strong persistence of DI dependence. Health indicators follow closely: the number of days hospitalized and the number of hospitalizations, together with normalized wage in the previous year, stand out as highly predictive. These measures distinguish between long hospital stays and repeated admissions, and between workers with stable labour market attachment and those with weaker earnings records. Other healthcare utilization measures, such as visits to general practitioners and to specialist doctors, also rank high, confirming that medical history is a powerful determinant of DI entry. Among sociodemographic variables, age is the most informative, but it becomes relevant only after these health and labour market factors, suggesting that demographics alone are much less predictive once richer information is available.

A particularly important predictor is whether the individual has experienced a DI spell in the previous two years. This variable dominates the importance rankings and highlights the strong recurrence of DI dependence: having been on DI before greatly increases the likelihood of returning. Including this factor as a predictor is valuable, since it captures an essential dimension of risk related to persistence. At the same time, we are also interested in the determinants of first-time entry, where persistence cannot play a role. To that end, we re-estimate the model excluding individuals with a prior DI spell. Predictive performance remains high (r^2=0.29, AUC = 0.905), and variable-importance rankings are largely unchanged (Figure 6, Panel B). Wage and hospitalization continue to be the most informative predictors.

This stability indicates that even when restricting attention to new entrants, health and earnings histories are the central determinants of DI risk.

Robustness checks

- Repeat the model for other years and check consistency of AUC and R^2 (figure with the distribution of predicted probabilities for 2 years in the sample plot)
- Sample restrictions: repeat excluding youngers (<25) or older (>55) individuals, to prove results are not carried by any age bracket, same for income brackets and blue/white collar workers.
- Change some variable definitions (i.e. reimbursement variables for healthcare consumption, hospitalizations (days vs. stays) drugs (quantity vs. by ddd).
- Do a basic linear regression to compare whether key predictors are the same.

5.2. Transition risk from ST-DI to LT-DI

Long-term DI cases are particularly relevant because individuals have already remained in the program for at least one year, which inherently entails higher expenditure and reflects a more serious situation than short-term or temporary cases. Understanding who transitions from short- to long-term DI is therefore crucial. Our baseline model for this outcome achieves an R^2 of 22.9%, slightly lower than for DI entry but still substantial. This value shows that close to one quarter of the heterogeneity in the probability of becoming long-term disabled can be explained by observable characteristics, which is considerable given the complexity of the transition and the fact that it conditions on already being on short-term DI.

Figure 6, Panel C, reports variable-importance scores for the full model. The ranking of predictors differs markedly from that for DI entry. Age is now the dominant factor, consistent with the idea that older workers who enter DI are less likely to return to work and more likely to remain on benefits. The prominence of psychiatric consultation as a predictor highlights that mental health problems play a central role in persistence. Pharmaceutical expenditure also ranks highly, suggesting that overall health burden, regardless of drug type, is informative of long-term dependence. Finally, labour market attachment, proxied by whether the individual was working full- or part-time before entering DI, also contributes meaningfully, capturing how weaker attachment increases the likelihood of remaining on benefits. Taken together, these results underscore that the drivers of DI entry and those of persistence differ; health and labour market factors play distinct roles across outcomes.

The distribution of predicted probabilities in Figure 4, Panel B, presents a different pattern compared to the entry model. Since the analysis is restricted to individuals already on DI, the

transition to long-term status is less rare, which results in a wider spread of predicted risks. Whereas probabilities were almost entirely concentrated near zero in the entry case, here a noticeable fraction of individuals reach values above 0.5. This distribution conveys an important implication: once individuals are in DI, the heterogeneity in their chances of remaining becomes much more visible, with a clear separation between those very likely to exit and those very likely to persist. In other words, conditional on entering DI, the risk of long-term dependence is not only higher on average but also more sharply differentiated across individuals.

Extension: mental-health subsample

An additional extension will focus specifically on long-term DI cases where the reported underlying pathology is mental health related. This is possible only for LT-DI, since diagnostic information is collected at the moment of reclassification but not during ST-DI. Given that psychiatric consultation emerges as a leading predictor in the full sample, narrowing the analysis to individuals with a mental health diagnosis at entry into LT-DI will allow us to investigate which patterns of healthcare use and labour history best signal the risk of becoming long-term disabled due to mental health conditions. This exercise aims to shed light on early warning signs that are specific to mental health–related disability.

Robustness checks

(same as for outcome 1)

- Repeat the model for other years and check consistency of AUC and R^2
- Sample restrictions: repeat excluding youngers (<25) or older (>55) individuals, to prove results are not carried by any age bracket, same for income brackets and blue/white collar workers.
- Change some variable definitions (i.e. reimbursement variables for healthcare consumption, hospitalizations (days vs. stays) drugs (quantity vs. by ddd).
- Do a basic linear regression to compare whether key predictors are the same.

5.3. Predictability by elapsed time and by horizon

We next analyse how predictability evolves with the duration of a DI spell and with alternative forecasting horizons. Conditioning on elapsed time in ST-DI, we estimate the probability of transition to LT-DI after one, two, three and four quarters, using only the information available up to each point. If the distribution of predicted probabilities becomes more dispersed at later quarters, with a clearer separation between low- and high-risk individuals, this would indicate that dynamic selection amplifies observable heterogeneity as time passes. Such evidence would highlight the potential for early warning signals: the ability to identify high-risk cases increases as a DI spell progresses, even before the formal reclassification to LT-DI at 12 months.

We also extend the analysis to different forecasting horizons. In addition to the one-year baseline, we estimate models for two- and five-year risks of DI entry and of being in LT-DI. Longer horizons naturally imply higher event rates and therefore more mass in the right tail of the predicted risk distribution. Comparing distributions across horizons shows whether current predictors remain informative in the medium run and whether a small subset of individuals consistently emerges as high risk. These exercises provide complementary perspectives: elapsed-time models capture within-spell dynamics, while multi-horizon models reveal how predictability evolves as the forecasting window lengthens. Together they inform both monitoring of ongoing DI spells and forward-looking policy planning.

Steps:

5.3.1. Elapsed time (quarters):
- Define risk groups at quarters=1, 2, 3, 4 in ST-DI.
- Train the models to predict entry into LT-DI.
- Calculate R^2, AUC and probabilities distribution.

5.3.2. Time horizons (2y and 5y):
- Create the subsamples with individuals that have info for 2y and 5y.
- Train the models to predict entry into LT-DI.
- Calculate R^2, AUC and probabilities distribution.

We will obtain:

- Figure with 2 panels:
    - A: probabilities distribution for each number of elapsed quarters (in the sample plot).
    - B: same for 2y and 5y in the same plot.

  The figure will show how predictive ability changes as time passes within ST-DI and as the prediction horizon lengthens.

5.4. Cross year variation and business cycle

Finally, we examine the stability of predictive performance across years and its sensitivity to the business cycle. For each year between 2006 and 2019, we estimate baseline models and compute R^2 and AUC. We then apply models trained in one year to predict outcomes in other years, measuring how predictive power decays with temporal distance. A slow decay would indicate that heterogeneity in DI risks is persistent, while a rapid decline would suggest that risk factors are more transitory or context dependent.
This cross-year exercise also allows us to relate predictive power to aggregate conditions. If models trained in recession years provide higher explanatory power than those trained in

expansion years, it would suggest that adverse labour market conditions strengthen selection and make DI risks more predictable. A figure analogous to Mueller and Spinnewijn's Figure 6 will display cross-year R^2 as a function of the lag between training and evaluation years, providing a direct visual summary of the persistence of heterogeneity and of the role of macroeconomic fluctuations.

Steps:

1. Train and estimate each model for 2006 to 2017
2. Calculate R^2, AUC and probabilities distribution for each year.
3. Cross-year: apply the training model to different years.
4. Represent R^2 as a function of time distance (lag).

We will obtain:
A figure with all the years in X-axis and R-squared in the Y-axis, highlighting recession periods. That would be informative of how the predictability evolves.
Interpretation: if it falls a little → persistent heterogeneity; if it falls a lot → transient factors and dependence on the macro context.

6. Conclusions

This paper has applied machine learning techniques to rich administrative data in Belgium to study the predictability of Disability Insurance entry and long-term dependence. We document substantial heterogeneity in DI risks: observable characteristics account for about one third of the variation in entry and close to one quarter of the variation in persistence, while a considerable share remains driven by unobserved factors. Health-related variables emerge as the most informative predictors, followed by labour market histories, whereas demographics alone explain little of the variation. Importantly, the determinants of DI entry and of persistence on the scheme differ, with medical history and previous DI spells dominating entry, and age, mental health, and weaker labour market attachment playing a central role in long-term dependence.

Beyond these baseline results, ongoing work explores robustness across subsamples, the evolution of predictability with elapsed time and forecasting horizons, and its stability across years and business-cycle conditions. These extensions will further clarify whether heterogeneity in DI risks reflects persistent individual traits or transitory factors linked to the macroeconomic context. Taken together, our findings highlight both the potential and the limits of observable data in predicting DI outcomes, and open new perspectives for preventive policies aimed at identifying individuals most at risk of long-term dependence.

References

Alvarez, F., & Schimer, R. (2011). Search and rest unemployment. Econometrica, 79(1):75-122.

Autor, D., & Duggan, M. (2003). The rise in the disability rolls and the decline in unemployment. Quarterly Journal of Economics, 118(1), 157–205.

Bruyneel, L., Rygaert, X., Oslejova, J., Avalosse, H., Fabri, V., Noirhomme, C., Willaert, D., Vrancken, J., Meeus, A., Leclercq, A., Karakaya, G., Brunois, T., Di Zinno, T., & Roelants, E. (2024). Incapacité de travail de longue durée et invalidité dues à des troubles psychosociaux – Profil socio-démographique, médical et de consommation de soins (Rapport). Agence Intermutualiste – INAMI.

Carey, C., Miller, N.H., & Molitor, D. (2022). Why does disability increase during recessions? Evidence from Medicare. NBER Working Papers 29988, National Bureau of Economic Research, Inc.

Charles, K.K., Li, Y., & Stephens, M. (2018). Disability benefit take-up and local labor market conditions. The Review of Economics and Statistics, MIT Press, vol. 100(3), 416-423.

Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants. A causal Machine Learning evaluation of training in Belgium". Labour Economics, 80, 102306.

De Brouwer, O., & Tojerow, I. (2023). The growth of disability insurance in Belgium: Determinants and policy implications (IZA Discussion Paper No. 16376). IZA – Institute of Labor Economics.

French, E., & Song, J. (2014). The effect of disability insurance receipt on labor supply. American Economic Journal: Economic Policy, 6(2), 291–337.

Holmgren, K., Fjällström-Lundgren, M., & Hensing, G. (2013). Early identification of work-related stress predicted sickness absence in employed women with musculoskeletal or mental disorders: A prospective, longitudinal study in a primary health care setting. Disability and Rehabilitation, 35(5), 418–426.

Institut national d'assurance maladie-invalidité (INAMI). (2018). Facteurs explicatifs relatifs à l'augmentation du nombre d'invalides : Régime des salariés et régime des indépendants. Période 2007–2016 [Study]. Brussels: INAMI.

Institut national d'assurance maladie-invalidité (INAMI). (2023). Incapacités de travail en 2023 : combien d'invalidités en raison d'une dépression ou d'un burnout ? Quel coût pour l'assurance indemnités ? INAMI. https://www.inami.fgov.be/fr/statistiques/statistiques-indemnites/statistiques-sur-les-incapacites-de-travail-decoulant-d-un-burnout-ou-d-une-depression/incapacites-de-travail-en-2023-combien-d-invalidites-en-raison-d-une-depression-ou-d-un-burnout-quel-cout-pour-l-assurance-indemnites

Kostøl, A. R., & Mogstad, M. (2014). How financial incentives induce disability insurance recipients to return to work. American Economic Review, 104(2), 624–655.

Maestas, N., Mullen, K. J., & Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. American Economic Review, 103(5), 1797–1829.

Moreau, M., Valente, F., Mak, R., Pelfrene, E., De Smet, P., De Backer, G., & Kornider, M. (2004). Occupational stress and incidence of sick leave in the Belgian workforce: The Belstress study. Journal of Epidemiology & Community Health, 58(6), 507–516.

Toppinen-Tanner, S., Ojajärvi, A., Väänänen, A., Kalimo, R., & Jäppinen, P. (2005). Burnout as a predictor of medically certified sick-leave absences and their diagnosed causes. Behavioral Medicine, 31(1), 18–32.

FIGURES & TABLES

Table 1: Variables included in each model

Sociodemographic	Health	Pharma	District extension	Labor market	Income
Gender	Visits with GP	Drugs related to a mucusloskeletal disorder – bought in a pharmacy	District	Ever unemployed	Labor income
Age	Visits with specialist doctor	Drugs related to a mucusloskeletal disorder – administered in a public hospital		Working time	
Married	Visits with psychiatra or psychologist	Drugs related to the nervous system – bought in a pharmacy		Type of worker	
Kids	Visit with a Kine	Drugs related to the nervous system – administered in a public hospital		Ever self-employed	
Foreign Nationality	Number of hospitalizations	Antidepressants from a pharmacy		Previous DI spell	
Region	Days hospitalized	Antidepressants in a hospital			

Note: For historical information we use the two prior years. Labor income refers to the income received the prior year. Health and pharma variables are included in the following different forms: a dummy variable indicating at least one single observation and the number of times. Working time can be full or part-time, type of worker can be public worker, blue-collar employee or white-collar employee. In Belgium there are 3 regions and 43 districts.

Table 2: Descriptive Statistics

Population vs. DI Sample		
	Population	DI sample
Age (mean)	40.9	44.8
Female	50.1%	40.5%
Foreign	22.2%	22.5%
Ever "Group N" drug in pharma	34.4%	49.9%
Ever "Group N" drug in hospital	7.5%	13.7%
Ever visit with a GP	95.1%	99.6%
Ever hospitalized	51.6%	78.0%
Ever Unemployment spell	6.8%	11.4%
Ever DI spell	6.5%	100%
Ever transition to LT-DI	2.7%	7.4%

Note: Descriptive statistics for the baseline sample for the years 2006-2018 and ages 16-65. DI sample include individuals in either short or long-term DI.

Table 3: AUC and R^2 for various models in 2018

Probability to enter DI		
	AUC	R^2
Sociodemographics	0.65	0.025
Sociodemographics + Labor Market	0.79	0.088
Sociodemographics + Health	0.86	0.214
All predictors	0.91	0.316

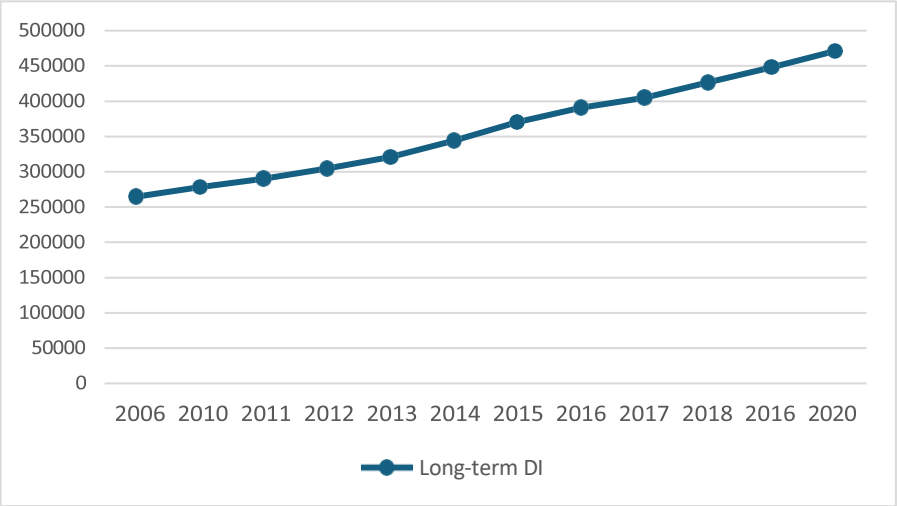
Note: AUC & R^2 for the ensemble model (RF + GBDT) using different sets of predictors for outcome 1 (probability to enter DI) for 2018.

Table 4: AUC and R^2 varying predictors in 2018 model.

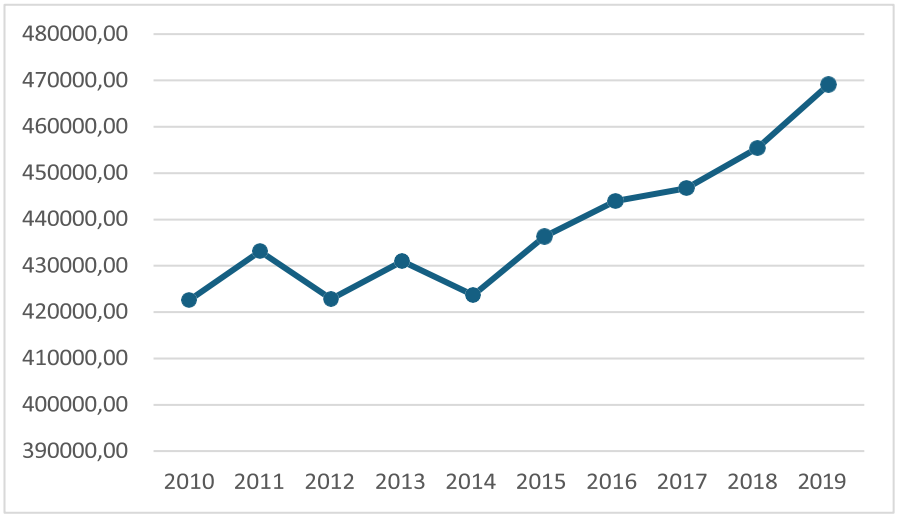
[work in progress]

Figure 1: Long and short-term DI evolution in Belgium

Panel A: Long-term DI cases

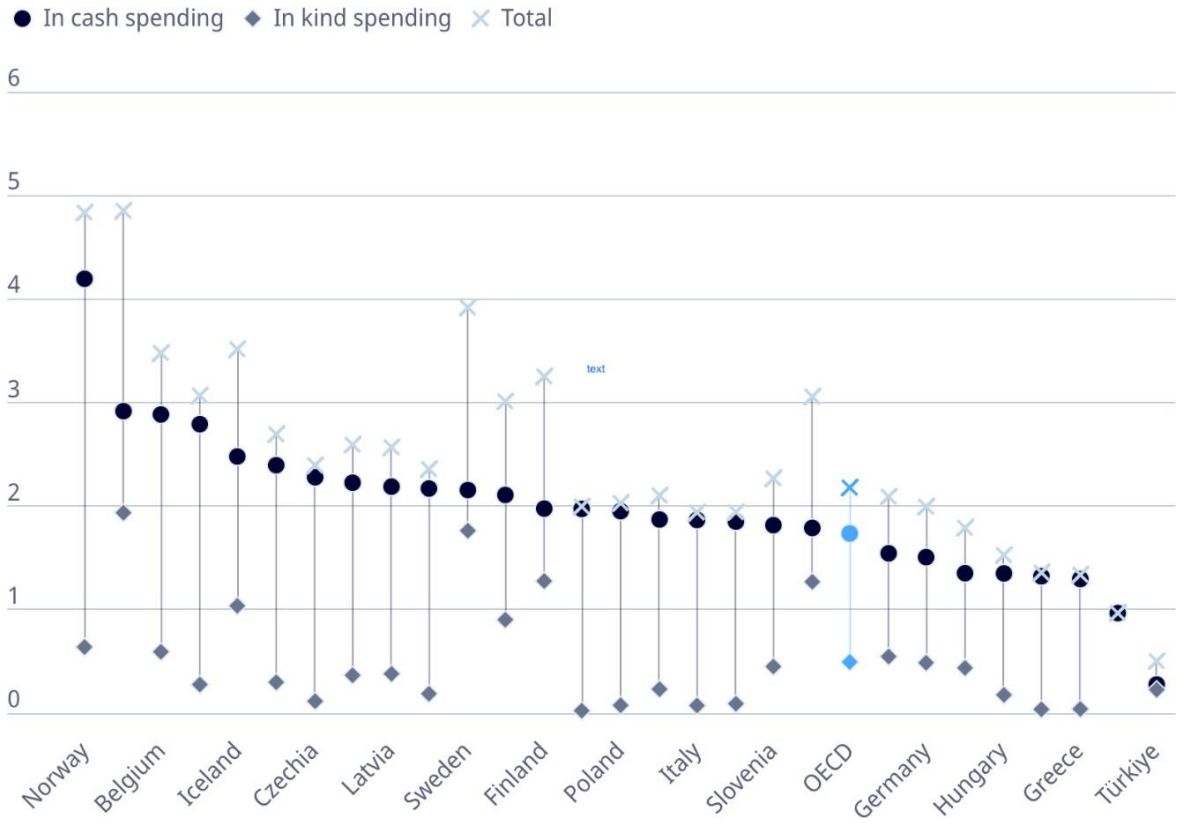


Panel B: Net entries on short-term DI



Note: self-constructed graphics with data from the National Institute for Health and Disability Insurance (NIHDI; INAMI in French). Panel A represent the total number of individuals on the long-term disability insurance program (more than 1 year) while panel B represents the net entries; the number of people starting a DI spell on that year.

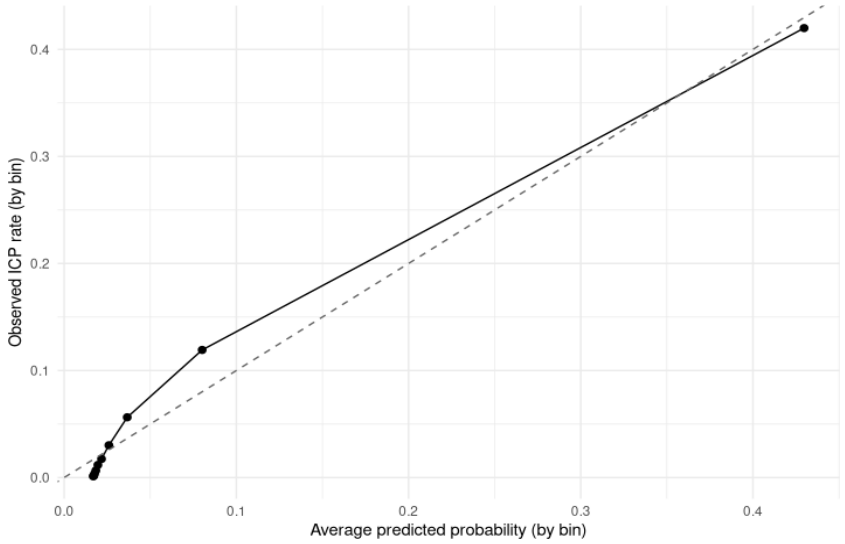
Figure 2: Public spending on incapacity in 2020



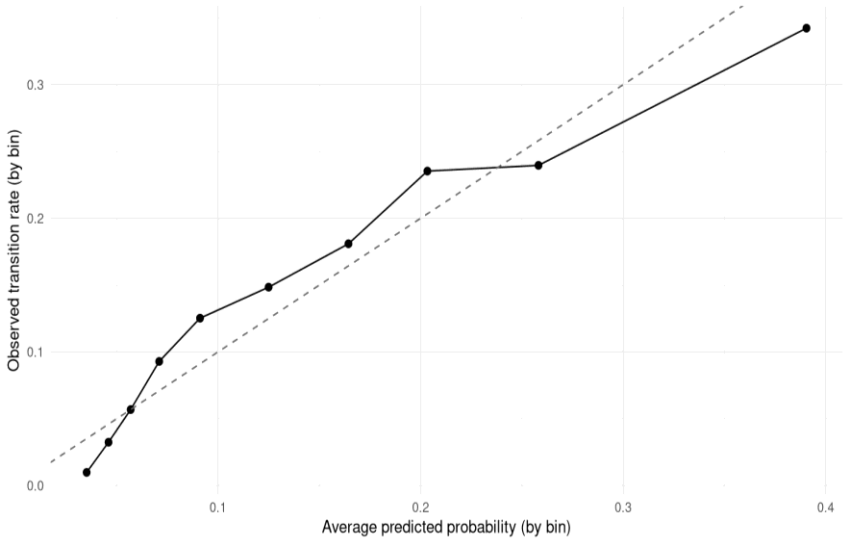
Note: Public spending on incapacity in % of GDP in the OCDE countries in 2020, it includes disability payments in cash, in kind and the sum of both. Data comes from OCDE statistics, 2020.

Figure 3: Comparing predictions from the baseline model to outcomes in 2018

Panel A: Outcome 1 (probability of entering DI)



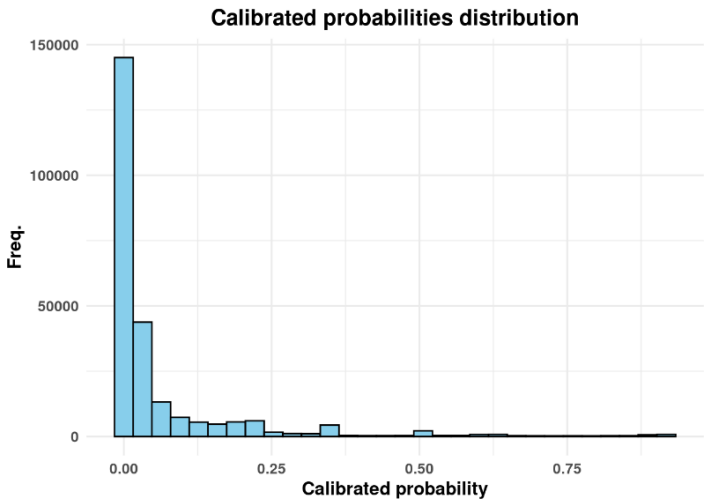
Panel B: Outcome 2 (probability of transitioning from ST- to LT-DI)



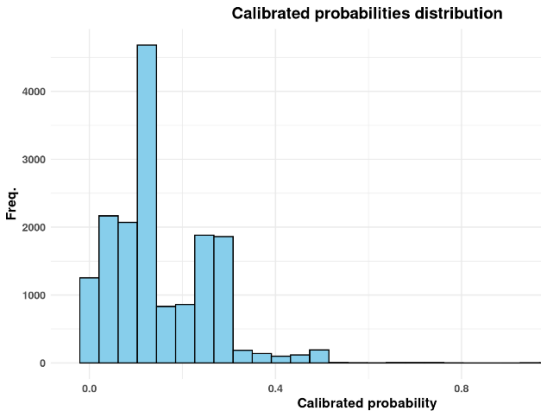
Note: Panel A shows a calibration plot comparing predicted and observed DI entry in the hold-out sample for 2018. Individuals are grouped into 10 deciles based on predicted DI entry probability. For each bin, we plot the average predicted probability against the observed DI entry rate. The dashed line represents perfect calibration (i.e., predicted = observed). Panel B presents the equivalent calibration plot for the probability of transitioning from short-term to long-term disability. As in Panel A, individuals are grouped into 10 bins of predicted probability, and the average observed transition rate is compared to the average predicted value in each bin.

Figure 4: Distribution of probabilities

Panel A: Outcome 1 (probability of entering DI)



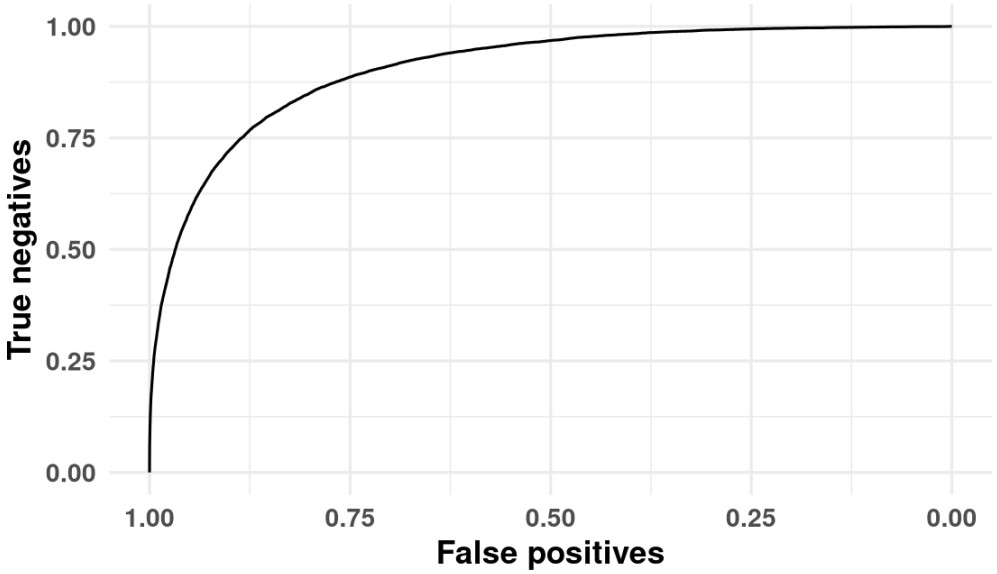
Panel B: Outcome 2 (probability of transitioning from ST- to LT-DI)



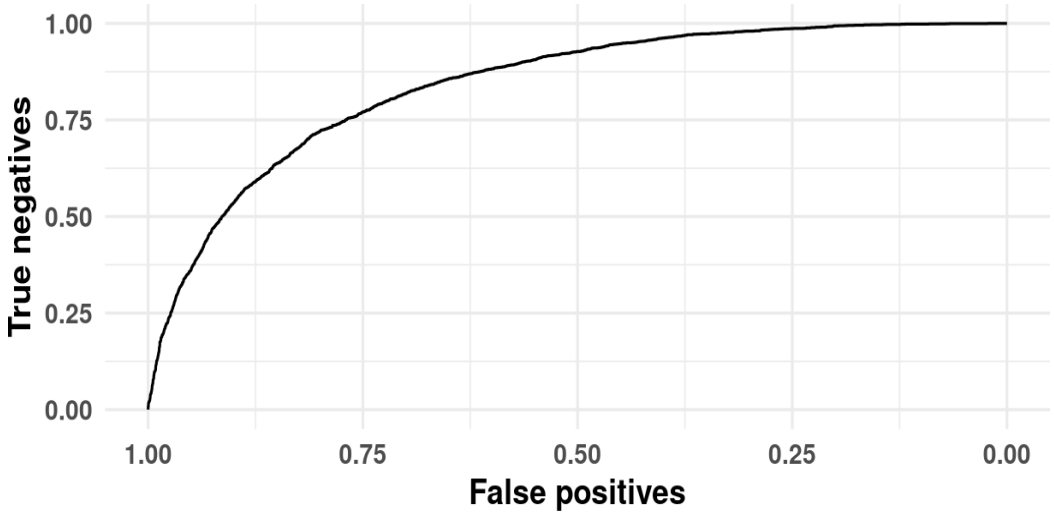
Note: Panel A shows a calibration plot comparing predicted and observed DI entry in the hold-out sample for 2018. Individuals are grouped into 10 deciles based on predicted DI entry probability. For each bin, we plot the average predicted probability against the observed DI entry rate. The dashed line represents perfect calibration (i.e., predicted = observed). Panel B presents the equivalent calibration plot for the probability of transitioning from short-term to long-term disability. As in Panel A, individuals are grouped into 10 bins of predicted probability, and the average observed transition rate is compared to the average predicted value in each bin.

Figure 5: ROC curves

Panel A: Outcome 1 (probability of entering DI)



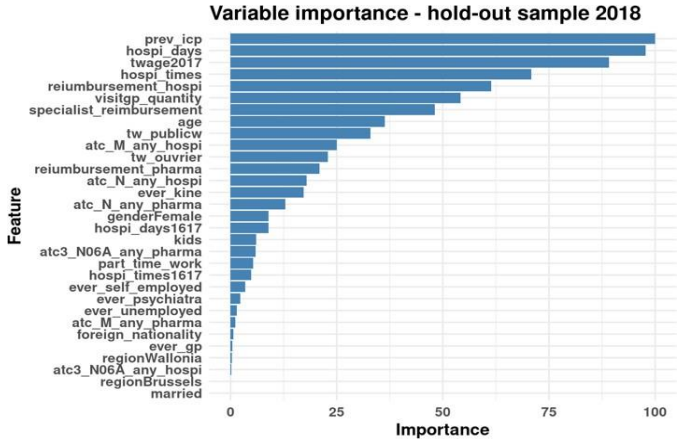
Panel B: Outcome 2 (probability of transitioning from ST- to LT-DI)



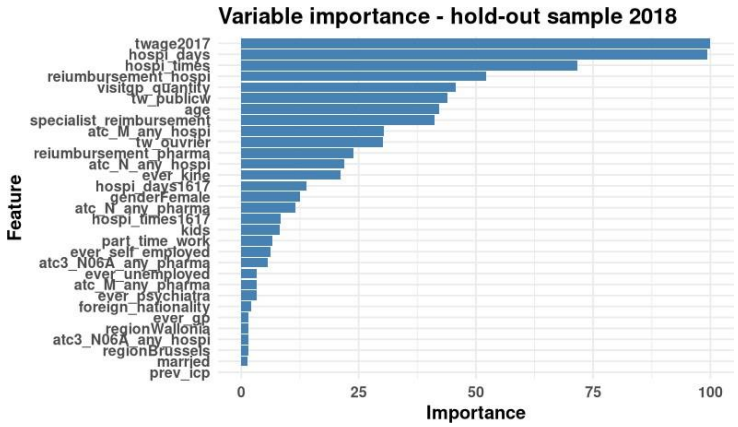
Note: Receiver operating characteristic (ROC) curves of the ensemble model for outcomes 1 & 2 evaluated in the hold-out sample for 2018.

Figure 6: Variables importance

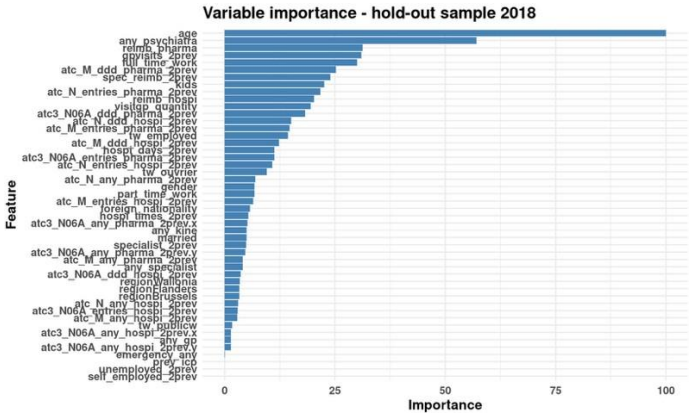
Panel A: Outcome 1 (probability of entering DI)



Panel B: Outcome 1 (excluding individuals with a previous DI spell)



Panel C: Outcome 2 (probability of transitioning from ST- to LT-DI)



Note: Contribution of each variable to the predictive power of the three different ensemble models for 2018.