Voorspellen van de instroom in de invaliditeitsverzekering en langdurige arbeidsongeschiktheid: Een benadering via Machine Learning

* Dit rapport is opgesteld voor het RIZIV

Guida Ayza Estopà (guida.ayza.estopa@ulb.be)
& Ilan Tojerow (ilan.tojerow@ulb.be)
Novembre 2025

1. Inleiding

Arbeidsongeschiktheid (AO) vormt een essentiële pijler van de sociale bescherming in ontwikkelde economieën. In de afgelopen decennia is het aantal arbeidsongeschikten aanzienlijk toegenomen in de OESO-landen, wat zorgen baart over de factoren die deze groei veroorzaken. Terwijl sommige landen erin geslaagd zijn het aantal arbeidsongeschikten de laatste jaren te stabiliseren of zelfs te verminderen, heeft België daarentegen een opvallende stijging gekend. Het aandeel van de bevolking op arbeidsleeftijd dat een invaliditeitsuitkering ontvangt (d.w.z. langer dan een jaar arbeidsongeschikt is) is tussen 2005 en 2020 bijna verdubbeld, van 3,5% naar 6,8% (RIZIV, 2022). Tegelijkertijd heeft België ook een sterke toename van de jaarlijkse instroom naar primaire arbeidsongeschiktheid gekend, vooral vanaf 2015. Deze dubbele evolutie plaatst het land bij de OESO-landen met de hoogste overheidsuitgaven voor arbeidsongeschiktheid. België heeft zelfs het op één na hoogste niveau van monetaire uitgaven voor arbeidsongeschiktheid. Rekening houdend met zowel geldelijke als niet-geldelijke prestaties, liggen de uitgaven net achter die van Scandinavische landen zoals Noorwegen, Denemarken en Zweden. In 2020 vertegenwoordigden ze 3,5% van het BBP (OESO, 2025) (zie Figuur A1 in de Bijlage voor een illustratie van deze trends).

Aangezien AO een cruciale rol speelt bij het beschermen van individuen die geconfronteerd worden met langdurige gezondheidsproblemen, is het begrijpen van de mechanismen achter de snelle groei van het aantal arbeidsongeschikten een centraal vraagstuk geworden, gezien de belangrijke maatschappelijke gevolgen. Op gezondheidsvlak kan dit immers wijzen op een verslechtering van het fysieke of mentale welzijn van een deel van de bevolking. Op de arbeidsmarkt gaat het gepaard met een lagere participatie en een grotere afstand tot werk. Ten slotte oefent langdurige arbeidsongeschiktheid vanuit financieel oogpunt een toenemende druk uit op de overheidsuitgaven en roept het vragen op over de houdbaarheid op lange termijn van het systeem. In deze context zijn politieke debatten steeds meer gericht op

preventieve strategieën, die erop gericht zijn het aantal nieuwe arbeidsongeshcikten te verminderen en hun determinanten beter te begrijpen. Sommige studies hebben al geprobeerd de aanhoudende stijging van het aantal invaliditeitsuitkeringen in de afgelopen jaren te verklaren. Autor en Duggan (2006) benadrukken dat deze groei niet uitsluitend kan worden toegeschreven aan de evolutie van de gezondheidstoestand van de bevolking of demografische factoren, maar dat het ook institutionele, economische en incentive-gerelateerde veranderingen weerspiegelt. In dezelfde lijn tonen De Brouwer, Leduc en Tojerow (2019) aan dat de versterking van de verplichtingen om werk te zoeken sommige individuen ertoe heeft gebracht zich tot invaliditeitsregelingen te wenden, wat heeft

bijgedragen aan een deel van de waargenomen stijging. Ondanks deze bijdragen blijft een belangrijk deel van de dynamiek echter nog onvoldoende begrepen, zoals ook Saks (2017) en De Brouwer & Tojerow (2023) suggereren.

Andere studies wijzen op meer specifieke mechanismen om de instroom in AO of de overgang naar invaliditeit te verklaren. Verschillende onderzoeken benadrukken met name de rol van financiële prikkels (Maestas et al., 2013; Kostol & Mogstad, 2021; Marie & Vall-Castelló, 2023; Ayza et al., 2025), de impact van institutionele hervormingen of de interactie met andere sociale programma's (Campolieti & Riddel, 2012; Fontenay & Tojerow, 2025; De Brouwer et al., 2023), evenals — in mindere mate — genderverschillen (Low & Pistaferri, 2019), administratieve vertragingen (Autor et al., 2015) of informatiefricties (Kostol & Myhre, 2021). Meer recent zijn ook de kwaliteit van werk en psychosociale risico's geïdentificeerd als opkomende determinanten van AO. Het geval van burn-out illustreert deze trend goed: aanvragen op basis van deze reden zijn de laatste jaren sterk toegenomen, en verschillende studies bevestigen bovendien dat slechte arbeidsomstandigheden of hoge werkstress belangrijke voorspellers zijn van instroom in AO (Clumeck et al., 2009; Holmgren et al., 2012; Nekoei et al., 2025).

Met focus op België suggereert Saks (2017) dat de recente stijging van het aantal arbeidsongeschikten deels voortkomt uit een grotere participatie aan de arbeidsmarkte van vrouwen en oudere werknemers. Echter, De Brouwer en Tojerow (2023) tonen via een meer uitgebreide analyse, die meerdere dimensies in rekening brengt, aan dat veranderingen in observeerbare kenmerken zoals geslacht, leeftijdsstructuur of jobkenmerken slechts een marginale rol spelen in de stijging van instroom in invaliditeit.

Een recente studie voor België (RIZIV & IMA, 2023) past voorspellende modellen toe om socio-demografische, medische en zorggerelateerde factoren te identificeren die het risico op invaliditeit bij personen met mentale gezondheidsproblemen kunnen aangeven. Hun resultaten tonen aan dat het gebruik van medische diensten, in het bijzonder psychiatrische consultaties en medicatie voor het zenuwstelsel, helpt om de overgang naar invaliditeit te voorspellen, hoewel het discriminerend vermogen van het model beperkt is. Hoewel dit een veelbelovende vooruitgang is, berust hun aanpak op een logistische regressie, waardoor ze zich beperken tot individuen met mentale stoornissen en een kleiner aantal voorspellers gebruiken. Daarentegen past dit artikel meer flexibele Machine Learning-methoden toe, bouwt het krachtigere voorspellende modellen en bestrijkt het de volledige bevolking op arbeidsleeftijd.

Dit artikel heeft twee hoofddoelstellingen. Ten eerste wil het begrijpen wat de determinanten zijn van de toename van het verzekerden in primaire arbeidsongeschiktheid. Ten tweede beoogt het de ontwikkeling van effectievere beleidsmaatregelen op het gebied van preventie en begeleiding te ondersteunen: enerzijds door vooraf de individuen te identificeren die het

meest waarschijnlijk arbeidsongeschikt worden, en anderzijds, onder degenen die zich al in het systeem bevinden, door degenen te herkennen die meer kans hebben om erin te blijven of, omgekeerd, eruit te stappen. Een betere identificatie van deze profielen zou potentieel kunnen leiden tot beleidsmaatregelen die nauwkeuriger zijn afgestemd op de kenmerken en behoeften van de betrokken personen.

Om dit te bereiken, maakt dit artikel gebruik van ML-technieken om te identificeren hoe verschillende sets van geobserveerde factoren (beroepshistoriek, gezondheidsgerelateerde determinanten, sociaaleconomische kenmerken) en niet-geobserveerde factoren interageren om de trajecten binnen AO vorm te geven, en zo de meest relevante voorspellers van het AO-risico te bepalen. Het gebruik van ML stelt ons in staat deze complexe interacties vast te leggen, de betrouwbaarheid van voorspellingen te verbeteren en een flexibele, datagestuurde benadering te hanteren om de dynamiek van AO te begrijpen. Onze analyse is gebaseerd op het kader van Mueller & Spinnewijn (2023), die de dynamiek van de werkloosheidsverzekering onderzoeken vanuit het perspectief van heterogeniteit, dynamische selectie en duur van de arbeidsongeschiktheid. Wij breiden hun kader uit naar AO, door te verkennen hoe vooraf bestaande individuele verschillen, de evolutie van de risicogroep in de tijd en duur-effecten zowel de instroom in primaire arbeidsongeschiktheid als de overgang naar invaliditeit beïnvloeden.

1.1. Gegevens, methodologie en bijdrage

In dit artikel worden drie complementaire modellen ontwikkeld om de dynamiek van arbeidsongeschiktheid te analyseren. Het eerste model voorspelt de kans om in een bepaald jaar in primaire arbeidsongeschiktheid terecht te komen. Het tweede model richt zich op personen die al primair arbeidsongeschikt zijn en schat hun kans om over te gaan naar invaliditeit. Het derde model onderzoekt ten slotte de kans om het systeem te verlaten tijdens het eerste jaar van arbeidsongeschiktheid, om zo het belang van dynamische selectie en duur van de arbeidsongeschiktheid te meten.

Deze drie oefeningen maken het mogelijk om verschillende centrale vragen te onderzoeken:

- In welke mate bestaat er heterogeniteit in de risico's op instroom in primaire arbeidsongeschiktheid, overgang naar invaliditeit en uitstroom?
- Welke dimensies (gezondheid, beroepsverleden, sociaaleconomische kenmerken) zijn het meest voorspellend?
- Domineren gezondheidsvariabelen, zoals ziekenhuisopnames, medicijngebruik of frequentie van medische consultaties, of spelen beroepsloopbanen een vergelijkbare rol?

- Ten slotte, wat betreft de overgang naar invaliditeit: laten bepaalde kenmerken toe om individuen te identificeren die bijzonder vatbaar zijn om langdurig in het uitkeringsstelsel te blijven?

Het begrijpen van deze dynamieken is essentieel, zowel om doeltreffendere sociale verzekeringssystemen te ontwerpen als om de voorspellingsinstrumenten voor het risico op primaire arbeidsongeschiktheid en invaliditeit te verbeteren.

De analyse steunt op een rijke dataset die administratieve gegevens over loopbanen, het gebruik van gezondheidszorg en sociodemografische kenmerken combineert, afkomstig uit twee hoofdbronnen: de Kruispuntbank van de Sociale Zekerheid (KSZ) en het Intermutualistisch Agentschap (IMA) in België. De dataset bestrijkt de periode 2006–2019 en vertegenwoordigt 10% van de Belgische bevolking, ofwel 735.000 individuen met kwartaalobservaties, wat neerkomt op bijna 40 miljoen observaties. Tabel 1 geeft een overzicht van de verschillende variabelensets die voor de analyse zijn gebruikt, en Tabel 2 bevat de beschrijvende statistieken voor de volledige steekproef en voor de substeekproef van AO (arbeidsongeschiktheid), bestaande uit individuen die een periode van AO hebben gekend.

Het artikel steunt op een conceptueel kader dat illustreert hoe verschillende bronnen van heterogeniteit, geobserveerd en niet-geobserveerd, de instroom in AO en de kans om langdurig in het stelsel te blijven kunnen beïnvloeden. De instroom in AO en de duur van arbeidsongeschikhteid worden gevormd door dynamieken die vergelijkbaar zijn met die in de werkloosheidsverzekering, met name door heterogeniteit van het initiële risico, dynamische selectie en de duur van arbeidsongeschiktheidDe heterogeniteit van het AO-risico vloeit voort uit individuele verschillen in observeerbare kenmerken zoals leeftijd, geslacht, inkomen, arbeidsomstandigheden of gezondheidstoestand (weergegeven via verschillende dimensies van gezondheidsuitgaven), evenals uit niet-geobserveerde kenmerken zoals veerkracht of de neiging om zorg te zoeken. Dynamische selectie treedt op omdat personen met hogere gezondheidsrisico's of een zwakkere arbeidsmarktbinding minder geneigd zijn om uit arbeidsongeschiktheid te stappen, waardoor de samenstelling van de groep die in AO blijft geleidelijk verandert. Dit proces is vergelijkbaar met de dynamieken in de werkloosheidsverzekering, waar de meest inzetbare personen sneller uitstappen, al kan het hier complexer zijn. Technische Bijlage C biedt een gedetailleerde bespreking van dit kader. Terwijl een uitgebreide literatuur al een aanzienlijke heterogeniteit in terugkeer naar werk bij werklozen heeft gedocumenteerd (Álvarez & Shimer, 2011;

Cockx et al., 2023), bestaat er weinig informatie over de omvang van de heterogeniteit en de implicaties ervan voor arbeidsongeschiktheid-uitkomsten.

Table 1 : Variabelen opgenomen in elk model

Sociodémographique	Marché du travail	Hospitalisation	Médecins	Pharma
Genre	A été au chômage	Jours d'hospitalisation	Consultations avec un médecin généraliste	Médicaments liés à un trouble musculosquelettique – achetés en pharmacie
Âge	Temps de travail	Nombre d'hospitalisation	Consultations avec un médecin spécialiste	Médicaments liés à un trouble musculosquelettique – administrés dans un hôpital public
Marié	Type de travailleur	Remboursement lié à l'hospitalisation	Consultations avec un psychiatre ou un psychologue	Médicaments liés au système nerveux – achetés en pharmacie
Enfants	A été indépendant	Passages aux urgences	Consultations avec un kiné	Médicaments liés au système nerveux – administrés dans un hôpital public
Nationalité Étrangère	Épisodes antérieurs d'incapacité du travail		Remboursements des consultations avec un médecin spécialiste	Antidépresseurs achetés en pharmacie
Région	Salaire			Antidépresseurs administrés dans un hôpital
				Remboursements des dépenses pharmaceutiques

Note : Voor de historische informatie worden de twee voorgaande jaren gebruikt. Het arbeidsinkomen verwijst naar het inkomen dat in het voorgaande jaar werd ontvangen. De medische consultaties en de farmaceutische variabelen zijn opgenomen in verschillende vormen: een indicatorvariabele die aangeeft of er minstens één observatie is, en het totale aantal voorkomens. De arbeidstijd kan voltijds of deeltijds zijn; het type werknemer kan een werknemer in de publieke sector zijn, een arbeider ("Blue Collar") of een bediende ("White Collar").

Al deze elementen ondersteunen het gebruik van ML-modellen, die in staat zijn complexe en niet-lineaire interacties vast te leggen, en tonen hun voordelen ten opzichte van traditionele econometrische benaderingen voor het voorspellen van arbeidsongeschiktheid. Voor de empirische implementatie wordt er gebruik gemaakt van de van standaard ML-technieken. De voorspellende modellen werden getraind op een trainingssteekproef en de prestaties

werden geëvalueerd op een validatiesteekproef om overfitting te vermijden. De belangrijkste uitdaging bij alle voorspellingsopdrachten ligt in het compromis tussen het verbeteren van het model en het risico op overfitting wanneer te veel variabelen worden opgenomen. ML-methoden en de scheiding van steekproeven helpen bij het optimaliseren van de variabelenselectie en het beheren van dit compromis tussen voorspellingsnauwkeurigheid en overfitting in een data-intensieve omgeving.

Er wordt gefocust op drie hoofdresultaten: (i) de kans om in primaire arbeidsongeschiktheid terecht te komen, (ii) de kans om van primaire arbeidsongeschiktheid naar invaliditeit over te gaan, en (iii) de kans om tijdens het eerste jaar uit arbeidsongeschiktheid te treden. Deze kansen worden gedefinieerd als de risicovariabelen van het model.

Er worden twee machine learning-modellen gebruikt: Random Forest en Gradient Boosted Regression Trees, respectievelijk gebaseerd op de aggregatie van onafhankelijke bomen en op een sequentiële constructie die opeenvolgende fouten corrigeert, waardoor complexe interacties en niet-lineariteiten kunnen worden vastgelegd. Deze worden vervolgens samengevoegd in een Ensemble-model, dat een gewogen lineaire combinatie van beide is. Nadat de modellen zijn afgestemd op de trainingssteekproef, worden ze ze geschat om de voorspellingen van het Ensemble-model en de gekalibreerde kansen voor elk resultaat te verkrijgen, en wordt het getrainde model toegepast op een controlegroep ("Hold-out sample") die niet eerder is gebruikt.

Table 2 : Beschrijvende statistieken

Échantillon de population vs. Échantillon IT		
	L'ensemble de l' échantillon	Échantillon IT
Âge (moyenne)	40.9	44.8
Femme	50.1%	40.5%
Étranger	22.2%	22.5%
Médicament "Group N" en pharmacie	34.4%	49.9%
Médicament "Group N" en hôpital	7.5%	13.7%
Médecin généraliste	95.1%	99.6%
Hospitalisation	51.6%	78.0%
Épisode de chômage	6.8%	11.4%
Épisode d'incapacité travail antérieur	6.5%	100%
Épisode d'invalidité antérieur	2.7%	7.4%

Nota : Beschrijvende statistieken voor de basissteekproef die de jaren 2006–2018 en de leeftijdscategorie 16–65 jaar bestrijkt. De AO-steekproef omvat individuen met kortdurende of langdurige arbeidsongeschiktheid..

Dit artikel draagt op drie hoofdmanieren bij aan de literatuur. Ten eerste biedt het een eerste systematische documentatie van de heterogeniteit van risico om in te stromen in

arbeidsongeschiktheid en tijdens de transities naar langdurige arbeidsongeschiktheid. Ten tweede, door ML toe te passen op rijke administratieve gegevens, kwantificeert het het aandeel van heterogeniteit dat toe te schrijven is aan observeerbare factoren en legt het complexe interacties vast. Ten derde levert het nieuwe bewijzen over de dynamische dimensie van het risico op arbeidsongeschiktheid, waarbij wordt aangetoond hoe de voorspelbaarheid evolueert met de duur van de periodes en de voorspellingen op langere termijn Samen openen deze conclusies nieuwe perspectieven voor het ontwerp van preventieve beleidsmaatregelen door de profielen te identificeren die het grootste risico lopen om in AO terecht te komen en er langdurig in te blijven.

2. Resultaten

Nu worden de resultaten van de drie voorspellende modellen gepresenteerd. Ten eerste wordt het basismodel onderzocht dat de kans schat om in primaire arbeidsongeschiktheid terecht te komen. Ten tweede wordt het model dat de transities naar invaliditeit voorspelt geanalyseerd. Ten derde worden de dynamieken tijdens het eerste jaar van arbeidsongeschiktheid bestudeerd.

2.1. Kans op het begin van een periode van arbeidsongeschiktheid

Het eerste model voorspelt de kans dat een individu in een bepaald jaar in primaire arbeidsongeschiktheid terechtkomt, op basis van informatie uit de twee voorgaande jaren. De basisspecificatie richt zich op voorspellingen voor 2018 en is geschat op een willekeurige steekproef van 10% van de Belgische bevolking op arbeidsleeftijd die in de twee voorgaande jaren geen enkele periode van arbeidsongeschiktheid heeft gekend. Voor elk individu kent het model een voorspelde kans toe om een nieuwe periode van primaire arbeidsongeschiktheid te beginnen.

Om de kwaliteit van een voorspellend model te beoordelen, worden twee hoofdmaatstaven gebruikt: de AUC en de R^2. De AUC (Area Under the Curve) meet het vermogen van het model om individuen die arbeidsongeschikt worden correct te onderscheiden van degenen die dat niet worden; een waarde van 0,5 komt overeen met toeval, terwijl een waarde dicht bij 1 een hoge discriminatie weerspiegelt. In ons geval bereikt de AUC 0,86, wat aangeeft dat het model zeer effectief onderscheid maakt tussen individuen met een hoog risico en die met een laag risico – een opmerkelijk resultaat voor een zeldzaam en multifactorieel fenomeen. De R^2 meet op zijn beurt het aandeel van de variatie in instroom in arbeidsongeschiktheid dat kan worden verklaard door observeerbare kenmerken. Met een R^2 van 18,9% is ongeveer een vijfde van de variatie toe te schrijven aan factoren zoals gezondheid, inkomen of werkhistoriek. Hoewel dit niveau in absolute termen bescheiden lijkt, is het hoog voor dit type voorspellende oefening. Ter vergelijking: Mueller en Spinnewijn (2023) behalen een R^2 van ongeveer 15% bij het

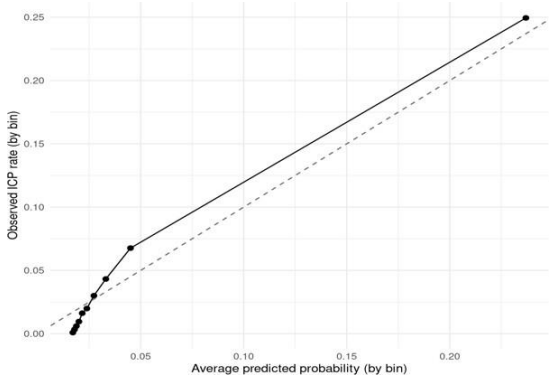
voorspellen van uitstroom uit de werkloosheidsverzekering met vergelijkbare administratieve gegevens.

Panel A van Figuur 1 vergelijkt de voorspelde kansen met de feitelijke instroom in primaire arbeidsongeschiktheid in 2018 en toont een duidelijke overeenkomst: groepen met een hoger voorspeld risico vertonen hogere waargenomen instroompercentages. Deze nauwe overeenstemming tussen voorspellingen en realisaties illustreert het sterke discriminerende vermogen van het model en benadrukt de kwaliteit van de verkregen voorspelling. Panel B toont de verdeling van de voorspelde kansen in de actieve bevolking. Zoals verwacht concentreert het merendeel van de individuen zich rond nul, aangezien de meerderheid nooit een periode van arbeidsongeschiktheid kent, terwijl een kleine minderheid duidelijk opvalt met hogere voorspelde risico's.

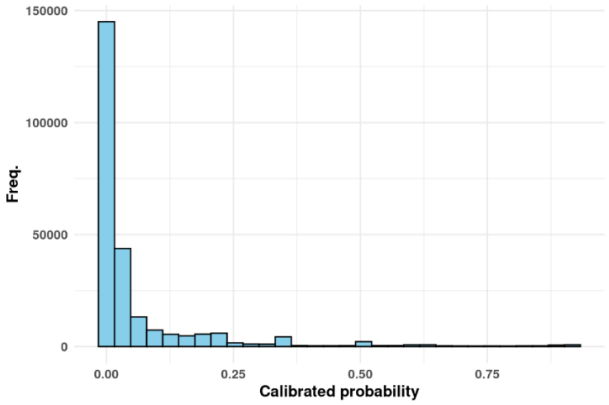
Figuur A3 in de bijlage toont de verdeling van de voorspelde kansen na uitsluiting van alle waarden onder 0,05 om het bovenste deel van de verdeling beter zichtbaar te maken. Zelfs in deze ingekorte steekproef blijven de meeste individuen geconcentreerd op relatief lage risiconiveaus, wat de zeldzaamheid van instroom in primaire arbeidsongeschiktheid weerspiegelt. Het model kent echter kansen toe over het volledige bereik tot waarden dicht bij één. Hoewel deze hoog-risicogevallen zeer zeldzaam zijn, tonen ze aan dat het model erin slaagt een kleine groep individuen met zeer hoge voorspelde risico's te identificeren. De verdeling blijft sterk asymmetrisch, wat onthult dat een kleine groep het grootste deel van het voorspelbare risico concentreert. Dit patroon benadrukt het potentieel voor vroege identificatie en gerichte preventie.

Figuur 1 : Kans op het begin van arbeidsongeschiktheid in 2018

Panel A: Vergelijking tussen de voorspellingen van het model en de waargenomen resultaten



Panel B: Verdeling van de voorspelde kans op arbeidsongeschiktheid



Nota : Panel A toont een calibratiegrafiek waarin de voorspelde en waargenomen kansen op instroom in arbeidsongeschiktheid worden vergeleken in de validatiesteekproef voor het jaar 2018. De individuen zijn gegroepeerd in 10 decielen op basis van hun voorspelde kans op instroom in arbeidsongeschiktheid. Voor elke groep geven we de gemiddelde voorspelde kans en het waargenomen instroompercentage weer. De stippellijn vertegenwoordigt perfecte calibratie (d.w.z. voorspelling = observatie). Omdat de grafiek gemiddelden per deciel gebruikt, dekken de waarden niet het volledige interval 0–1, hoewel de individuele voorspellingen dat wel doen. Panel B toont de verdeling van de voorspelde kans op instroom in arbeidsongeschiktheid in de validatiesteekproef voor het jaar 2018.

Om te begrijpen welke factoren het meest bijdragen aan de voorspellende nauwkeurigheid, toont Tabel 3 de prestaties van modellen die gebaseerd zijn op verschillende groepen variabelen. Panel A voegt geleidelijk informatie over de arbeidsmarkt en gezondheid toe aan een basismodel dat enkel socio-demografische kenmerken bevat. Alleen socio-demografische kenmerken verklaren slechts een klein deel van de variatie in het instroomrisico. Het toevoegen van arbeidsmarktinformatie verhoogt het verklarend vermogen met 167% ten opzichte van het basismodel. Het opnemen van gezondheidsvariabelen zorgt voor de grootste verbetering: het verklarend vermogen stijgt dan met 278%. Dit bevestigt dat gezondheidsindicatoren veruit de krachtigste voorspellers zijn van het risico op arbeidsongeschiktheid.

Panel B onderzoekt de marginale bijdrage van de verschillende subgroepen van gezondheidsvariabelen: ziekenhuisopnames dragen het meest bij, gevolgd door het geneesmiddelengebruik, terwijl consultaties bij specialisten of huisartsen bescheidener maar significante verbeteringen toevoegen.

Table 3 : R^2 voor verschillende modellen in 2018

A. Submodellen van het referentiemodel – Sequentieel

	c (1)	c (2)	c (3)
R^2	0.0187	0.0499	0.1887
Variation (c) vs (c-*1*)		+167%	+278%
Sociodémographique	X	X	X
Variables du marché du travail		X	X
Variables liées à la santé			X

B. Uitsplitsing van de bijdrage van de verschillende gezondheidsdimensies – Marginaal

	c (1)	c (2)	c (4)	c (5)	c (3)	c (6)
R^2	0.0187	0.0238	0.0281	0.0335	0.1128	0.1301
Variation (c) vs (*1*)		+27%	+50%	+79%	+503%	+596%
Sociodémographique	X	X	X	X	X	X
Médecin généraliste		X				X
Spécialistes			X			X
Consommation des médicaments				X		X
Hospitalisations					X	X

Nota : De tabel toont de R^2 die wordt verkregen door de voorspelde kans op arbeidsongeschiktheid te regresseren op een variabele die de daadwerkelijke arbeidsongeschiktheid in de validatiesteekproef voor het jaar 2018 aangeeft. Panel A begint met het basismodel in c (1) en voegt de groepen variabelen sequentieel toe totdat alle variabelen van het referentiemodel in c (3) zijn opgenomen. Panel B gebruikt dezelfde groepen variabelen, maar splitst de gezondheidsgerelateerde informatie op in verschillende subgroepen, die eerst afzonderlijk en vervolgens gelijktijdig worden toegevoegd in kolom c (6).

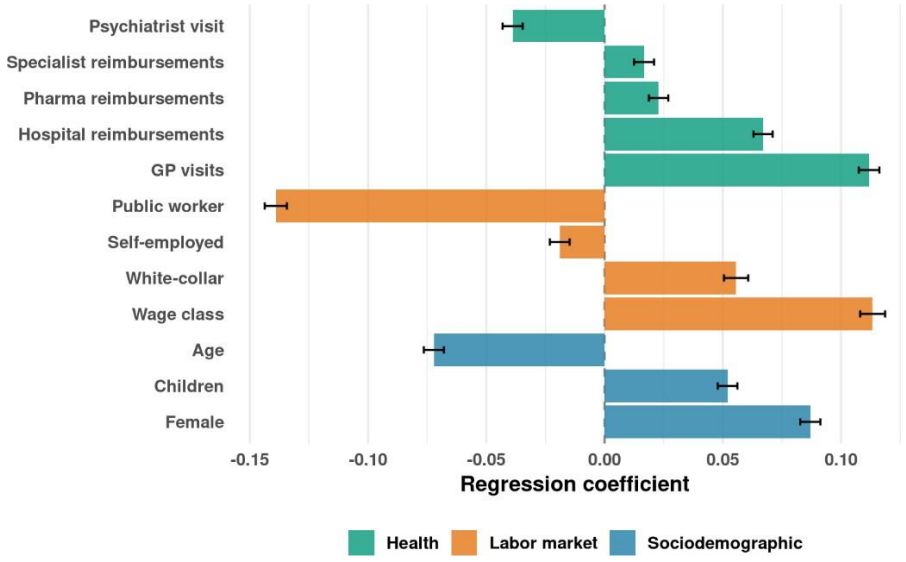
Hoewel het voorspellende model geen causale effecten schat, maakt het wel mogelijk om de bijdrage van verschillende groepen variabelen aan de voorspellende prestaties te evalueren. Onder de gezondheidsgerelateerde voorspellers komen het aantal ziekenhuisdagen en opnames naar voren als de belangrijkste, gevolgd door consultaties bij de huisarts en specialist. Bij de variabelen van de arbeidsmarkt is de loonklasse van het voorgaande jaar de sterkste voorspeller. Leeftijd is het meest relevante socio-demografische kenmerk zodra met andere informatie rekening wordt gehouden.

Om de relevantie van de variabelen op een intuïtieve manier te illustreren, werd er een een gestandaardiseerde lineaire regressie geschat van de gekalibreerde kansen op alle voorspellers, volgens Mueller en Spinnewijn (2023). De coëfficiënten worden uitgedrukt in standaardafwijkingen en tonen hoe elk kenmerk (positief of negatief) samenhangt met het voorspelde risico om in arbeidsongeschiktheid terecht te komen. Figuur 2 presenteert deze resultaten. Zoals verwacht domineren de gezondheidsvariabelen (in groen), met name medische consultaties, ziekenhuisopnames, farmaceutische terugbetalingen en consultaties bij specialisten. Interessant is dat psychiatrische consultaties een negatieve associatie vertonen, wat suggereert dat een vroege psychiatrische behandeling de kans op arbeidsongeschiktheid kan verminderen. De arbeidsmarktkenmerken (in oranje), zoals loonklasse en type werk, spelen ook een belangrijke rol. Positieve coëfficiënten voor hogere lonen en 'witteboorden'-beroepen kunnen erop wijzen dat personen met stabielere of beter beschermde banen minder beperkingen ervaren wanneer zij ziekteverlof nodig hebben. Daarentegen vertonen werknemers in de publieke sector en zelfstandigen negatieve associaties, wat erop wijst dat zij minder geneigd zijn een periode van arbeidsongeschiktheid te beginnen. De socio-demografische variabelen (in blauw) spelen een beperktere rol zodra met andere dimensies rekening wordt gehouden: vrouw zijn en kinderen hebben hangen samen met een hoger risico, terwijl leeftijd een negatieve associatie vertoont.

Het is belangrijk te benadrukken dat deze interpretatieve regressie niet weerspiegelt hoe de variabelen daadwerkelijk bijdragen aan de voorspellingen in het machine-learningmodel. In een niet-lineair model hangt de bijdrage van een variabele immers af van zowel haar interacties met andere kenmerken als van complexe relaties die een lineaire regressie niet kan vastleggen. De OLS-coëfficiënten vatten enkel gemiddelde en lineaire associaties samen, terwijl het ML-model veel rijkere verbanden in rekening brengt. Om die reden kunnen sommige categorische

variabelen, zoals het feit dat men een publieke werknemer is, hoge lineaire coëfficiënten vertonen ondanks een beperkte bijdrage aan de globale prestaties van het niet-lineaire model. We hebben ook het model geschat inclusief individuen die al een periode van arbeidsongeschiktheid hebben gekend. Dit model vertoont nog betere prestaties, met een AUC van 0,88 en een R^2 van 24,8%. Niet verrassend is de belangrijkste voorspeller in deze specificatie het feit dat men eerder een periode van arbeidsongeschiktheid heeft meegemaakt. Deze variabele vangt op zichzelf een groot deel van het voorspelbare risico, aangezien individuen die in de twee voorgaande jaren arbeidsongeschikt waren veel meer kans hebben op een nieuwe periode van arbeidsongeschiktheid. Hoewel deze persistentie informatief is over herhaling en langdurige arbeidsongeschiktheid, beperkt ze onze mogelijkheid om de eerste instroom te bestuderen. De resultaten van dit model verschijnen in Figuur A4 in de Technische Bijlage.

Figuur 2 : Heterogeniteit van het risico op arbeidsongeschiktheid



Nota : Deze figuur toont de resultaten van lineaire regressies van de voorspellingen op een subset van observeerbare variabelen. De variabelen zijn gestandaardiseerd door het gemiddelde van de steekproef af te trekken en te delen door de standaardafwijking, zodat de coëfficiënten kunnen worden geïnterpreteerd als de verandering (in standaardafwijkingen) van het resultaat die overeenkomt met een verandering van één standaardafwijking in de covariabele. Ze toont de OLS-coëfficiënten van een regressie van de voorspelde kans op instroom in arbeidsongeschiktheid (basismodel 2018) op de variabelen die op de y-as staan. Rond de puntschattingen staan betrouwbaarheidsintervallen van 95%. Arbeiders (blue collar) vormen de referentiecategorie voor het type werknemer; de coëfficiënten voor bedienden (witte boorden) en werknemers in de publieke sector worden dus relatief ten opzichte van deze categorie geïnterpreteerd. Voor indicatorvariabelen — vrouw, zelfstandige, kinderen hebben, bezoek aan een psychiater — is de weggelaten categorie "niet over deze eigenschap beschikken". De continue variabelen (leeftijd, loon, medische terugbetalingen) worden in gestandaardiseerde vorm opgenomen.

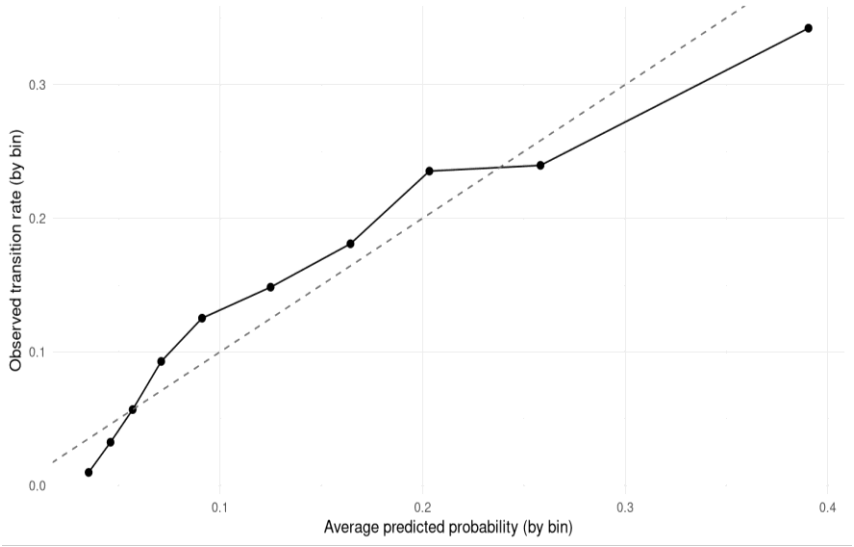
Over het geheel genomen tonen deze resultaten aan dat het model dat de kans voorspelt om in primaire arbeidsongeschiktheid terecht te komen een sterke voorspellende capaciteit en een duidelijke observeerbare heterogeniteit vertoont. De gezondheidstoestand en medische voorgeschiedenis zijn de belangrijkste bepalende factoren voor het risico op arbeidsongeschiktheid, gevolgd door factoren die verband houden met de arbeidsmarkt,

terwijl socio-demografische kenmerken relatief weinig bijdragen. Tegelijkertijd blijft ongeveer twee derde van de variantie onverklaard, wat suggereert dat niet-geobserveerde heterogeniteit — zoals niet-gemeten gezondheidscondities, veerkracht of individueel gedrag — een belangrijke rol blijft spelen.

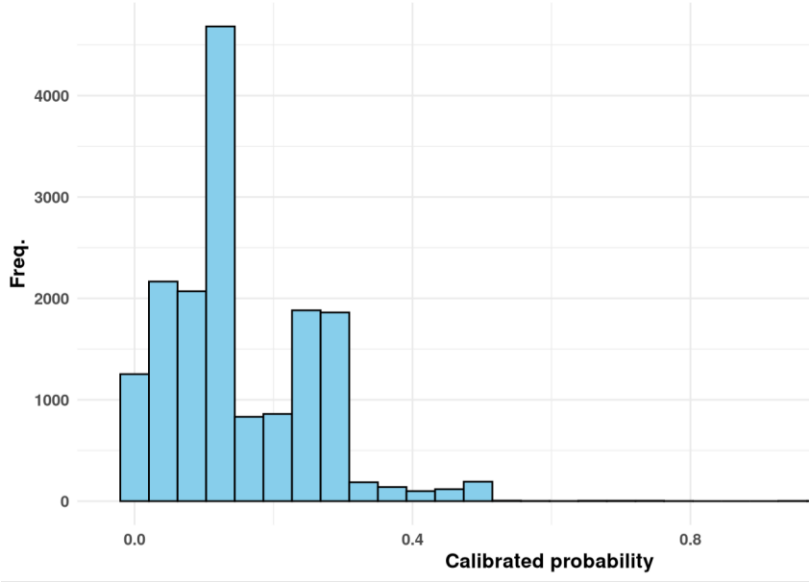
Kortom, de instroom in arbeidsongeschiktheid is sterk geconcentreerd bij een relatief kleine groep werknemers van wie de recente gezondheids- en werkhistoriek al een hoog risico signaleert. Deze concentratie biedt belangrijke kansen voor gerichte preventiebeleid en vroege interventie.

Figuur 3 : Kans op overgang van arbeidsongeschiktheid naar invaliditeit in 2018

Panel A : Vergelijking tussen de voorspellingen van het model en de waargenomen resultaten



Panel B : Verdeling van de voorspelde overgangskans



Nota : Paeel A toont een calibratiegrafiek waarin de voorspelde en waargenomen kansen op de overgang van arbeidsongeschiktheid naar invaliditeit worden vergeleken in de validatiesteekproef voor 2018. De individuen zijn gegroepeerd in 10 decielen op basis van hun voorspelde kans op instroom in arbeidsongeschiktheid. Voor elke groep geven we de gemiddelde voorspelde kans en het waargenomen overgangspercentage weer. De stippellijn vertegenwoordigt perfecte calibratie. Net zoals bij Figuur 1 dekken de waarden niet het volledige interval 0–1, omdat ze gebaseerd zijn op gemiddelden per deciel. Paneel B toont de verdeling van de voorspelde kans op overgang naar invaliditeit in de validatiesteekproef voor 2018.

2.2. Waarschijnlijkheid van overgang van arbeidsongeschiktheid naar invaliditeit

Ons tweede model richt zich op individuen die al arbeidsongeschikt zijn en voorspelt wie in de loop van het volgende jaar in invaliditeit zal terechtkomen. Na één jaar arbeidsongeschiktheid kunnen personen in aanmerking komen voor de invaliditeitsstatus, na een nieuwe medische evaluatie. Deze overgang brengt hen in het langdurige regime, dat minder medische herzieningen voorziet en specifieke regels toepast voor de berekening van uitkeringen. Deze regels blijven gebaseerd op het vorige loon en de samenstelling van het huishouden, maar met andere vervangingspercentages en minimum- en maximumdrempels, zodat het bedrag kan stijgen, dalen of gelijk blijven afhankelijk van de gezinssituatie en het loonniveau. Het begrijpen van deze overgang is cruciaal, omdat ze de persistentie van arbeidsongeschiktheid en de factoren die leiden tot langdurige arbeidsongeschiktheid weerspiegelt.

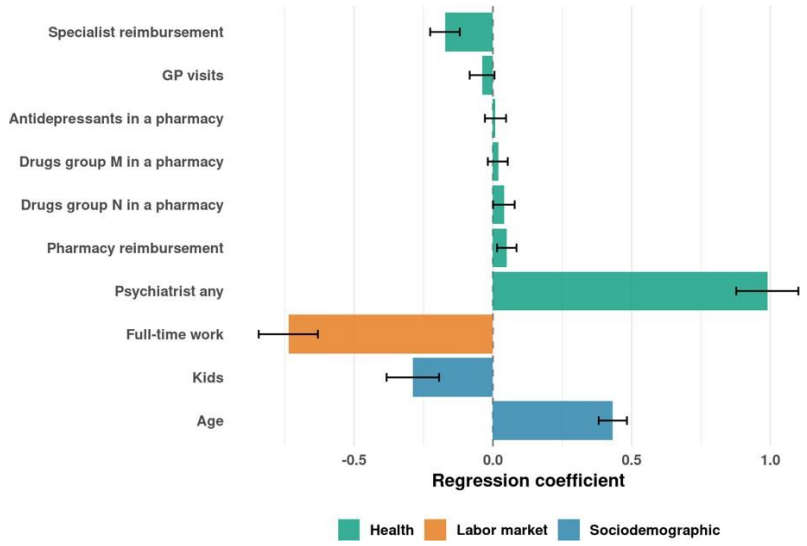
Het model bereikt een AUC van 0,72 en een R^2 van 0,073, wat aangeeft dat, hoewel de voorspellingen redelijk goed onderscheid maken tussen degenen die wel en niet naar invaliditeit overgaan, slechts een bescheiden deel van de totale variatie wordt verklaard door observeerbare kenmerken. Met andere woorden, het model kan individuen identificeren met een relatief hoger of lager risico, maar een groot deel van de persistentie lijkt te worden

bepaald door niet-geobserveerde factoren, zoals onderliggende gezondheidstrajecten, herstelprocessen of gedragingen die verband houden met moreel risico.

Panel A van Figuur 3 vergelijkt de waargenomen en voorspelde overgangskansen en toont dat de voorspelde risico's overeenkomen met de werkelijke resultaten, hoewel de spreiding groter is dan in het instroommodel. Panel B illustreert de verdeling van de voorspelde kansen. Omdat de analyse beperkt is tot individuen die al arbeidsongeschikt zijn, is de overgang naar invaliditeit minder zeldzaam dan de initiële instroom, wat leidt tot een meer uitgespreide verdeling van voorspelde risico's. Hoewel de meeste individuen zich op gemiddelde niveaus bevinden, overschrijdt een niet-verwaarloosbaar deel een voorspelde kans van 0,5, wat aangeeft dat het model een subgroep met hoog risico kan identificeren. Over het geheel genomen is de kans om van primaire arbeidsongeschiktheid naar invaliditeit over te gaan veel minder voorspelbaar dan de kans op een eerste instroom; niettemin blijft de analyse van de relevantie van voorspellers instructief ondanks de lagere verklarende kracht.

Figuur A5 in de Technische Bijlage toont de variabele-importance scores voor het volledige model, en Figuur 4 geeft de coëfficiënten van de gestandaardiseerde lineaire regressie weer, analoog aan de oefening voor het instroommodel. De rangschikking van voorspellers verschilt duidelijk van die in het instroommodel. Leeftijd blijkt de dominante factor en is geassocieerd met een positieve coëfficiënt, wat consistent is met het feit dat oudere werknemers minder herstelperspectieven hebben en meer kans maken om naar invaliditeit over te gaan. Indicatoren voor mentale gezondheid, zoals psychiatrische consultaties, behoren ook tot de belangrijkste voorspellers, wat suggereert dat psychische aandoeningen een centrale rol spelen in langdurige arbeidsongeschiktheid. Farmaceutische uitgaven zijn een andere belangrijke voorspeller, wat de algemene gezondheidsbelasting weerspiegelt in plaats van specifieke medische aandoeningen. Ten slotte is de arbeidsintensiteit vóór de arbeidsongeschiktheid (bijvoorbeeld voltijds of deeltijds werken vóór de instroom) eveneens van belang, wat erop wijst dat individuen met een zwakkere band met de arbeidsmarkt een hogere kans hebben om uit het werk te blijven.

Figuur 4 : Heterogeniteit van het risico op invaliditeit



Nota : Deze figuur toont de resultaten van lineaire regressies van de voorspellingen op een subset van observeerbare variabelen. De variabelen zijn gestandaardiseerd, waardoor de coëfficiënten kunnen worden geïnterpreteerd als het effect van een variatie van één standaardafwijking in de covariabele. Ze toont de OLS-coëfficiënten van de voorspelde kans op overgang naar invaliditeit (model 2018) voor de variabelen die op de y-as staan. Rond de schattingen staan betrouwbaarheidsintervallen van 95%. Voor indicatorvariabelen — voltijds werken, kinderen hebben, bezoek aan een psychiater — is de weggelaten categorie "niet over deze eigenschap beschikken". De continue variabelen (leeftijd, medicijngebruik, medische terugbetalingen) worden in gestandaardiseerde vorm opgenomen.

Over het geheel genomen benadrukken deze resultaten een duidelijke asymmetrie tussen de determinanten van de instroom in arbeidsongeschiktheid en die van de persistentie naar invaliditeit. De instroom is beter voorspelbaar op basis van vooraf bestaande gezondheidscondities en loopbaantrajecten, terwijl de persistentie na meer dan een jaar veel minder voorspelbaar is. Zodra individuen al arbeidsongeschikt zijn, spelen niet-geobserveerde gezondheids- of arbeidsdynamieken een steeds grotere rol. Twee complementaire mechanismen kunnen helpen om dit patroon te verklaren: *Dynamische selectie* — waarbij individuen met de beste herstelperspectieven vroeg uitstappen, waardoor de samenstelling van de resterende groep verandert; *Duur van de arbeidsongeschiktheid* — waarbij de kans om van primaire arbeidsongeschiktheid naar invaliditeit over te gaan toeneemt naarmate de tijd in arbeidsongeschiktheid langer wordt, als gevolg van een reële verslechtering of een aanpassing aan het systeem. We onderzoeken het relatieve belang van deze mechanismen in de volgende sectie door de evolutie van de voorspellende prestaties in de tijd te analyseren tijdens periodes van primaire arbeidsongeschiktheid.

2.3. Dynamiek van voorspelbaarheid binnen periodes van arbeidsongeschiktheid

Nu wordt de evolutie van de voorspellende prestaties tijdens periodes van arbeidsongeschiktheid bestudeerd, volgens de dynamische benadering van Mueller en Spinnewijn (2023). Deze analyse onderzoekt of het vermogen van het model om persistentie te voorspellen varieert naargelang de tijd die in arbeidsongeschiktheid wordt doorgebracht en de voorspellingen op langere termijn. Het begrijpen van deze dynamieken is essentieel om te bepalen of persistentie voornamelijk te wijten is aan dynamische selectie — waarbij individuen met hogere uitstroomkansen (betere gezondheid, betere inzetbaarheid, grotere veerkracht, …) vroeg arbeidsongeschiktheid verlaten, waardoor een residuele groep overblijft die homogener is en gekenmerkt wordt door lagere uitstroomkansen — of aan duur van de arbeidsongeschiktheid, waarbij de risico's evolueren op een manier die niet door observeerbare variabelen wordt vastgelegd, mogelijk gerelateerd aan moreel risico of een grotere vertrouwdheid met het systeem naarmate de tijd verstrijkt.

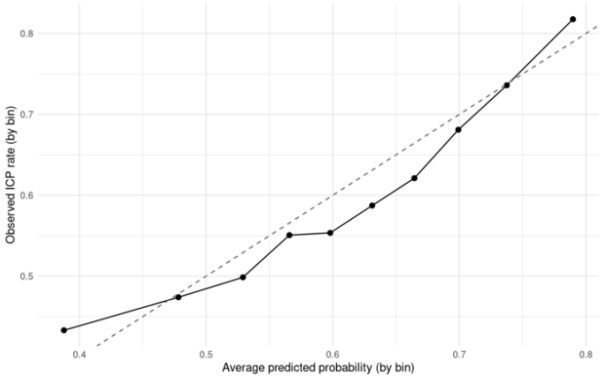
Er wordt gestart met een dynamisch referentiemodel dat voorspelt of individuen die net in arbeidsongeschiktheid zijn terechtgekomen minstens twee opeenvolgende kwartalen in arbeidsongeschikhteid blijven of eerder uitstappen, uitsluitend op basis van informatie die aan het begin van de periode beschikbaar is. Met andere woorden: of personen vóór of na twee kwartalen arbeidsongeschiktheid verlaten. Dit basismodel geeft weer in welke mate kortetermijnpersistentie vooraf kan worden voorspeld. Zoals bij de twee vorige resultaten vergelijkt panel A van Figuur 5 de waargenomen en voorspelde kansen, en toont dat de voorspelde risico's redelijk goed overeenkomen met de werkelijke resultaten, ook al is dit model duidelijk zwakker dan de vorige (AUC = 0,64; R^2 = 6%). Een AUC van 0,64 geeft aan dat het model mensen die arbeidsongeschiktheid verlaten beter onderscheidt van blijvers dan toeval, maar met een beperkte discriminerende capaciteit. Evenzo toont een R^2 van 6% dat slechts een klein deel van de waargenomen variatie in uitstroom wordt verklaard door de beschikbare variabelen, wat de intrinsieke moeilijkheid weerspiegelt om zeer korte trajecten aan het begin van de arbeidsongeschiktheid te voorspellen. Panel B toont de verdeling van de voorspelde kansen, die meer verspreid is en de grotere moeilijkheid weerspiegelt om kortetermijnpersistentie te onderscheiden bij individuen die net in arbeidsongeschiktheid zijn terechtgekomen.

Vervolgens wordt een conditioneel model geschat dat beperkt is tot individuen die twee kwartalen in arbeidsongeschiktheid blijven, en wordt er voorspeld of zij nog twee extra kwartalen zullen blijven of juist na vier kwartalen zullen uitstappen. De vergelijking van deze twee oefeningen biedt een directe evaluatie van de evolutie van het risico zodra de vroege uitstappen hebben plaatsgevonden. Panel A van Figuur 6 vergelijkt de verdelingen van voorspelde risico's aan het begin van de periode en na twee kwartalen. De verdeling in het tweede kwartaal is meer geconcentreerd en verschoven naar lagere uitstroomkansen, wat weerspiegelt dat individuen die na de eerste maanden in arbeidsongeschiktheid blijven, duurzaam lage uitstroomrisico's vertonen. Deze vermindering van de spreiding is consistent met dynamische selectie: individuen met hoog risico stappen vroeg uit, waardoor een

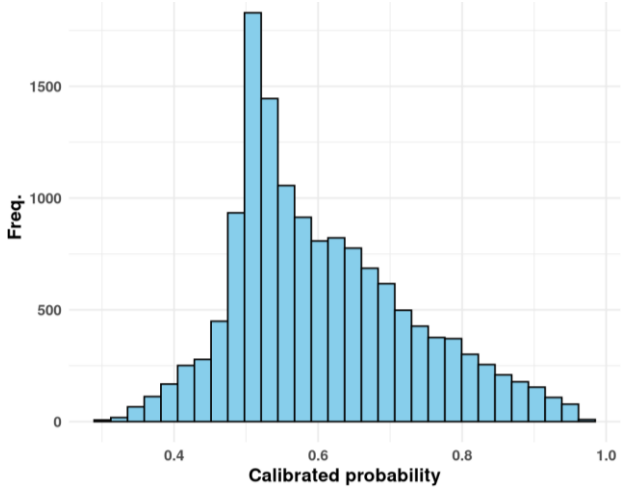
homogener groep overblijft met systematisch lagere kansen om arbeidsongeschiktheid te verlaten. Als gevolg daarvan verbetert de voorspellende prestatie licht, waarbij AUC en R^2 respectievelijk stijgen naar 0,65 en 8,5%. Deze waarden blijven bescheiden, maar tonen aan dat, zodra vroege uitstappen zijn geëlimineerd, observeerbare verschillen tussen individuen iets beter de uitstroomkansen verklaren. Als deze maten waren gedaald en als de verdeling was geconvergeerd naar een sterke persistentie ongeacht individuele kenmerken, zou dat op duur van de arbeidsongeschiktheid hebben gewezen.

Figuur 5 : Kans om in de eerste twee kwartalen van de periode (2018) weer aan het werk te gaan

Panel A : Vergelijking tussen de voorspellingen van het model en de waargenomen resultaten



Panel B : Verdeling van de voorspelde waarschijnlijkheid van uitval

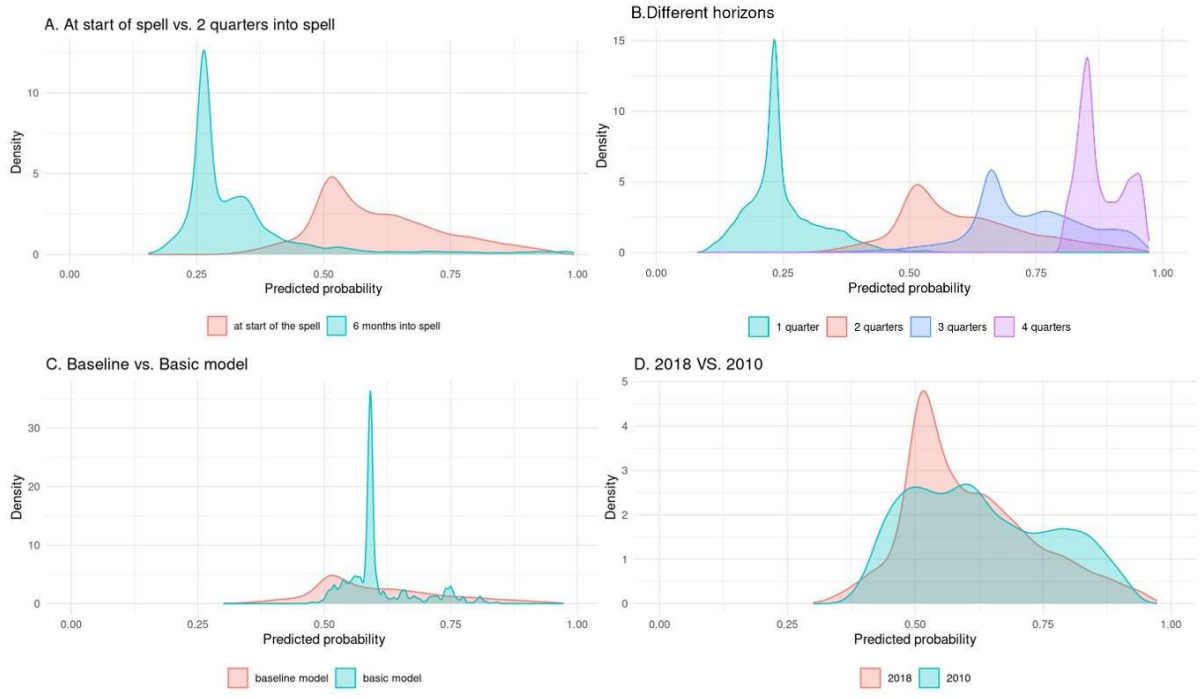


Note : Panel A toont een calibratiegrafiek waarin de voorspelde en waargenomen kansen op uitstroom uit arbeidsongeschiktheid tijdens de eerste twee kwartalen worden vergeleken in de validatiesteekproef voor 2018. De individuen zijn gegroepeerd in 10 decielen op basis van hun voorspelde kans op instroom in arbeidsongeschiktheid. Voor elke groep geven we de gemiddelde

voorspelde kans en het waargenomen uitstroompercentage weer. De stippellijn vertegenwoordigt perfecte calibratie. Paneel B toont de verdeling van de voorspelde kans op uitstroom in de validatiesteekproef voor 2018.

Vervolgens wordt er onderzocht hoe de voorspelbaarheid varieert naargelang de lengte van de voorspelling op lange termijn, terwijl de informatiebasis constant blijft. Panel B van Figuur 6 toont de verdelingen van voorspelde uitstroomrisico's voor toekomstige termijnen van één, twee, drie en vier kwartalen. De korte-termijnvoorspellingen zijn geconcentreerd rond lage uitstroomkansen, wat weerspiegelt dat slechts een klein deel van de individuen het regime in het volgende kwartaal verlaat. Naarmate de termijn zich uitstrekt tot twee of drie kwartalen, verschuiven de verdelingen naar rechts en worden ze meer verspreid, wat overeenkomt met de accumulatie van uitstroomrisico in de tijd en een grotere observeerbare heterogeniteit op middellange termijn. De voorspellingsnauwkeurigheid verbetert dienovereenkomstig: de AUC stijgt van 0,58 naar 0,64 en vervolgens 0,66, terwijl de R^2 toeneemt van 0,015 naar 0,056 en vervolgens 0,060. Wanneer de termijn echter vier kwartalen bereikt, dalen beide maten, wat suggereert dat lange-termijntrajecten steeds meer worden beïnvloed door niet-geobserveerde heterogeniteit en intrinsiek minder voorspelbaar worden. Tabel A1 (paneel B) in de bijlage vat de overeenkomstige waarden van AUC en R^2 samen.

Figuur 6 : Verdeling van de voorspelde kansen op het einde van de arbeidsongeschiktheid



Nota : Deze figuur toont de verdeling van verschillende voorspelde kansen op uitstroom uit arbeidsongeschiktheid. In de vier panelen komt de referentieverdeling overeen met de voorspelde uitstroomkansen over twee kwartalen, berekend bij het begin van de arbeidsongeschiktheidsperiode. Panel A vergelijkt deze referentieverdeling met de verdeling die wordt voorspeld voor de twee volgende kwartalen, onder de voorwaarde dat het individu minstens twee kwartalen in arbeidsongeschiktheid is gebleven. Panel B presenteert de voorspelde uitstroomkansen voor verschillende termijnen— één, twee, drie en vier kwartalen — op basis van hetzelfde informatiepakket. Panel C vergelijkt de voorspellingen van het volledige model (inclusief gedetailleerde variabelen over gezondheid, arbeidsmarkt en socio-demografie) met die van een gereduceerd model dat slechts een beperkt aantal

voorspellers gebruikt. Panel D vergelijkt de verdelingen van voorspelde risico's voor twee cohorten, 2010 en 2018, om de stabiliteit van de observeerbare heterogeniteit in de tijd te beoordelen.

Panel C beoordeelt in welke mate de voorspellende prestaties afhangen van de rijkdom van de informatiebasis. Een model dat beperkt is tot een klein aantal basisvariabelen genereert een verdeling met uitgesproken pieken, wat weerspiegelt dat het model met weinig voorspellers zeer vergelijkbare kansen toekent aan veel individuen. Daarentegen produceert het volledige model, dat gedetailleerde variabelen omvat met betrekking tot het gebruik van gezondheidszorg, loopbaantrajecten en socio-demografische kenmerken, gladdere en meer gedifferentieerde verdelingen. Deze vergelijking toont aan dat rijke administratieve informatie essentieel is voor relevante voorspellingen, en dat het grootste deel van de eerder gedocumenteerde voorspellende waarde rechtstreeks voortkomt uit de opname van gedetailleerde variabelen over gezondheid en arbeidsmarkt.

Ten slotte worden de verdelingen van voorspelde risico's voor twee cohorten vergeleken, 2010 en 2018, om na te gaan of de risicostructuur is geëvolueerd tijdens een periode waarin het aantal instromen in arbeidsongeschiktheid aanzienlijk is gestegen. De analyse van deze twee jaren maakt het mogelijk om te beoordelen in welke mate de risicostructuur in de tijd is veranderd. Beide cohorten vertonen een vergelijkbare algemene vorm, wat wijst op een relatieve stabiliteit van de observeerbare heterogeniteit die ten grondslag ligt aan de persistentie in invaliditeit gedurende het afgelopen decennium.

Zoals panel D van Figuur 6 laat zien, vertoont de cohorte van 2018 echter een veel uitgesprokener piek rond een kans van ongeveer 0,5, terwijl de verdeling van 2010 gladder en iets meer verspreid is. Deze sterkere concentratie in 2018 geeft aan dat een groter deel van de individuen zich bevindt rond een gemiddeld risico op persistentie, wat kan wijzen op een geleidelijke evolutie van de profielen van personen die in arbeidsongeschiktheid terechtkomen. Ondanks deze zichtbare vormverschillen blijft de globale structuur tussen de twee cohorten dicht bij elkaar, wat suggereert dat de heterogeniteitsmechanismen die door het model worden vastgelegd grotendeels stabiel blijven in de tijd.

Over het geheel genomen tonen deze analyses aan dat de persistentie in invaliditeit systematisch evolueert naarmate periodes vorderen. De voorspelbaarheid neigt toe te nemen na de eerste maanden, in overeenstemming met dynamische selectie onder vroege uitstappers, terwijl voorspellingen op langere termijnen een grotere niet-geobserveerde heterogeniteit onthullen. Rijke informatiebestanden verbeteren de discriminatie aanzienlijk, en de risicoprofielen lijken stabiel tussen cohorten. Samen benadrukken deze resultaten het belang van snelle en gedetailleerde informatie om periodes van arbeidsongeschiktheid op te volgen en individuen te identificeren van wie het risico op langdurige arbeidsongeschiktheid na verloop van tijd duidelijker wordt.

3. Conclusie

Dit artikel biedt een diepgaande analyse van de determinanten en de voorspelbaarheid van trajecten van arbeidsongeschiktheid in België, op basis van rijke administratieve gegevens en moderne machine-learningtechnieken. Drie hoofdbevindingen komen naar voren.

Ten eerste is dat de kans om in kortdurende arbeidsongeschiktheid terecht te komen sterk voorspelbaar is. Het basismodel vertoont een hoge discriminerende capaciteit en verklaart bijna een vijfde van de waargenomen variatie in instroomrisico's, een substantieel aandeel gezien de zeldzaamheid en complexiteit van het fenomeen. Gezondheidsindicatoren — in het bijzonder ziekenhuisopnames en farmaceutisch gebruik — zijn veruit de meest informatieve voorspellers, terwijl loopbaantrajecten een secundaire maar belangrijke rol spelen. Socio-demografische kenmerken dragen weinig bij zodra rekening wordt gehouden met gezondheids- en arbeidsmarktinformatie. De verdeling van voorspelde risico's is sterk asymmetrisch: de meeste individuen hebben een quasi nul-risico, terwijl een kleine groep het grootste deel van het voorspelbare risico concentreert. Dit patroon bevestigt het potentieel van gerichte preventieve interventies.

Ten tweede is de overgang van primaire arbeidsongeschiktheid naar invaliditeit aanzienlijk minder voorspelbaar. Hoewel het model individuen die wel of niet de overgang maken correct onderscheidt, wordt slechts een bescheiden deel van de variatie verklaard door observeerbare kenmerken. De rangschikking van voorspellers verandert ook: leeftijd wordt de belangrijkste factor, gevolgd door indicatoren voor mentale gezondheid en farmaceutische uitgaven, terwijl de binding met de arbeidsmarkt vóór de arbeidsongeschiktheid voorspellende kracht behoudt. Deze resultaten benadrukken een belangrijke asymmetrie: terwijl de instroom in arbeidsongeschiktheid grotendeels wordt bepaald door observeerbare gezondheids- en arbeidsprofielen, hangt de persistentie na het eerste jaar meer af van moeilijk meetbare factoren, zoals klinische evolutie, hersteltrajecten of een toenemende vertrouwdheid met het systeem van arbeidsongeschiktheid.

Ten derde onthult de dynamische analyse binnen periodes van arbeidsongeschiktheid dat dynamische selectie een centrale rol speelt in het vroege aanhouden binnen het regime. Individuen die in de eerste kwartalen in arbeidsongeschiktheid blijven, vertonen systematisch lagere uitstroomrisico's en worden een homogenere groep in termen van observeerbare voorspellers. Daardoor verbetert de voorspellende nauwkeurigheid tussen het eerste en het derde kwartaal. Daarentegen neemt de voorspelbaarheid af bij voorspellingen op langere termijn wat de toenemende invloed van niet-geobserveerde heterogeniteit weerspiegelt naarmate de tijd verstrijkt. Vergelijkingen tussen verschillende informatiebases bevestigen dat gedetailleerde gegevens over gezondheidszorg en loopbaantrajecten essentieel zijn voor

relevante voorspellingen, terwijl vereenvoudigde modellen er niet in slagen effectief te discrimineren. Ten slotte suggereert de stabiliteit van risicodistributies tussen de cohorten van 2010 en 2018 dat de observeerbare heterogeniteit die ten grondslag ligt aan persistentie in arbeidsongeschiktheid opmerkelijk stabiel blijft in de tijd en weinig wordt beïnvloed door macro-economische omstandigheden of institutionele hervormingen.

Gezamenlijk tonen deze resultaten de waarde van machine-learningbenaderingen om risicoprofielen te karakteriseren en het ontwerp van preventieve beleidsmaatregelen te sturen. Ze tonen aan dat arbeidsongeschiktheidsrisico's sterk geconcentreerd en in grote mate voorspelbaar zijn nog vóór het begin van een arbeidsongeschiktheidsperiode, wat duidelijke kansen biedt voor vroege identificatie en gerichte ondersteuning. Tegelijkertijd benadrukken de dynamieken van persistentie het belang van nauwgezette opvolging tijdens de eerste maanden van arbeidsongeschiktheid, wanneer individuele trajecten beginnen te divergeren. Deze inzichten kunnen bijdragen aan meer proactieve en beter gerichte interventies, wat uiteindelijk de beheersing van arbeidsongeschiktheidsrisico's en de duurzaamheid van het invaliditeitsverzekeringssysteem verbetert.

Referenties

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

Autor, D., & Duggan, M. (2006). The rise in the disability rolls and the decline in unemployment. Quarterly Journal of Economics, 118(1), 157–205.

Autor, D. H., Maestas, N., Mullen, K. J., & Strand, A. (2015). Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants. NBER Working Paper No. 20840, National Bureau of Economic Research.

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.

Campolieti, Michele & Riddell, Chris, 2012. "Disability policy and the labor market: Evidence from a natural experiment in Canada, 1998–2006," Journal of Public Economics, Elsevier, vol. 96(3), pages 306-316.

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.

Clumeck, N., Kempenaers, C., Godin, I., Dramaix, M., Kornider, M., Linkowski, P., & Kittel, F. (2009). Working conditions predict incidence of long-term spells of sick leave due to depression: results from the Belstress I prospective study. Journal of Epidemiology & Community Health, 63(4), 286–292.

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., Leduc, E., & Tojerow, I. (2023). The unexpected consequences of job search monitoring. Journal of Public Economics. 224: 104929

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.

Fontenay, S., & Tojerow, I. (2025). Is supported employment effective for Disability Insurance recipients with mental health conditions? Evidence from a randomized experiment in Belgium. Journal of Health Economics, 91, 103103.

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

Institut national d'assurance maladie-invalidité & Agence InterMutualiste (INAMI & AIM). (2024). Incapacité de travail de longue durée et invalidité dues à des troubles psychosociaux: Profil socio-démographique, médical et de consommation de soins. INAMI. https://aim-ima.be/IMG/pdf/rapport_aim_-_incapacite_de_travail_de_longue_duree_et_invalidite_dues_a_des_troubles_psychosociaux.pdf

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

Kostøl, A. R., & Myhre, A. S. (2021). Labor supply responses to learning the tax and benefit schedule. American Economic Review, 111(11), 3733–3766.

Low, H., & Pistaferri, L. (2019). Disability insurance: Error rates and gender differences. Journal of Political Economy, 127(6), 2839–2892.

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.

Marie, O,. & Vall Castelló, J. (2023). "Sick Leave Cuts and (Unhealthy) Returns to Work," Journal of Labor Economics, University of Chicago Press, vol. 41(4), pages 923-956.

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.

Nekoei, A., Salines, M., Schmieder, J., & von Wachter, T. (2025). Labor Market Effects of Social Insurance Reforms. CESifo Working Paper No. 11128.

Saks, Y. (2017). Better understanding the upward trend in the number of disability insurance claimants. National Bank of Belgium, issue ii, pages 55-68.

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.

Technische bijlage in het Engels bij het rapport : "Voorspellen van de instroom in de invaliditeitsverzekering en langdurige arbeidsongeschiktheid: Een benadering via Machine Learning"

Guida Ayza Estopà & Ilan Tojerow (DULBEA, ULB)

Novembre 2025

Technical appendix

This Technical Appendix accompanies the main French report for the study "Predicting Disability Insurance Entry and Long-Term Dependence: A Machine Learning Approach" and provides additional methodological detail, supplementary results, and a comprehensive description of the data and analytical framework used throughout the study. Its purpose is to offer full transparency on the empirical approach, the construction of variables, and the statistical models underlying the findings presented in the main document.

Using rich Belgian administrative data (2006–2019) and ensemble machine-learning models, we predict (i) entry into work incapacity, (ii) transitions to long-term disability, and (iii) early-spell exit dynamics. Results show that work-incapacity entry is highly predictable, mainly driven by health-related indicators, while transitions to long-term disability are less so and reflect greater unobserved heterogeneity. Dynamic analyses reveal strong early selection effects and broadly stable risk distributions across cohorts.

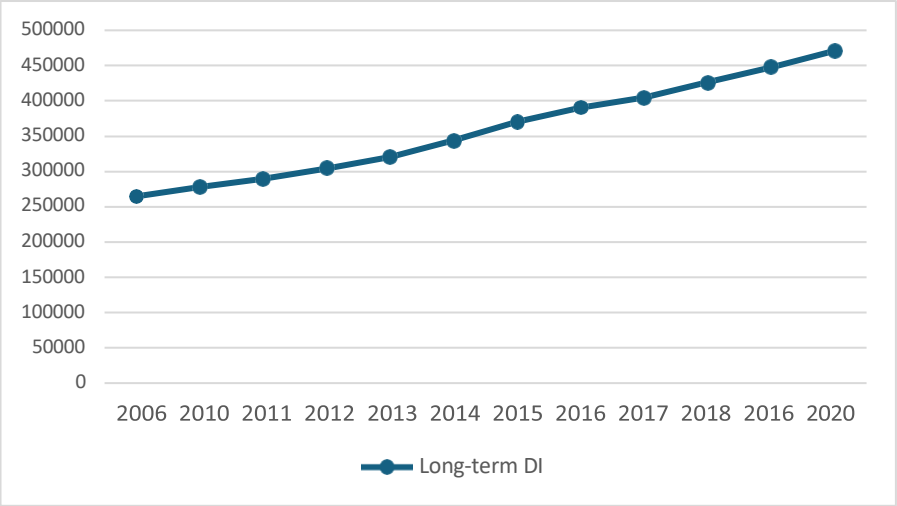
The appendix is structured into four parts.

Section A presents additional figures and tables that complement the descriptive evidence and empirical results discussed in the report.
Section B provides detailed information on the data sources, sample construction, variable definitions, and the institutional context of work incapacity and disability insurance in Belgium.
Section C outlines the conceptual framework used to interpret the dynamics of work incapacity, clarifying the role of heterogeneity, dynamic selection, and duration dependence.
Section D documents the prediction methodology in detail, including model training, variable selection, calibration procedures, and performance evaluation.

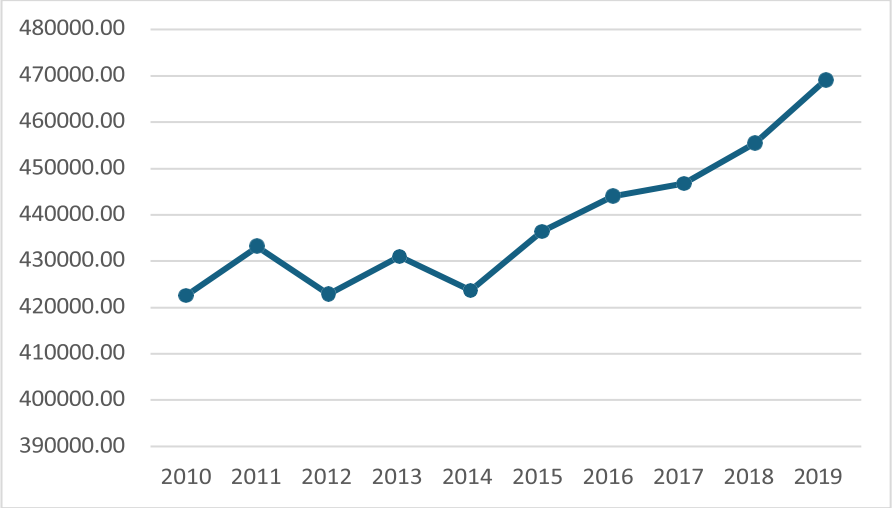
A. Additional Figures and Tables

Figure A1: Long and short-term DI evolution in Belgium

Panel A: Long-term DI cases

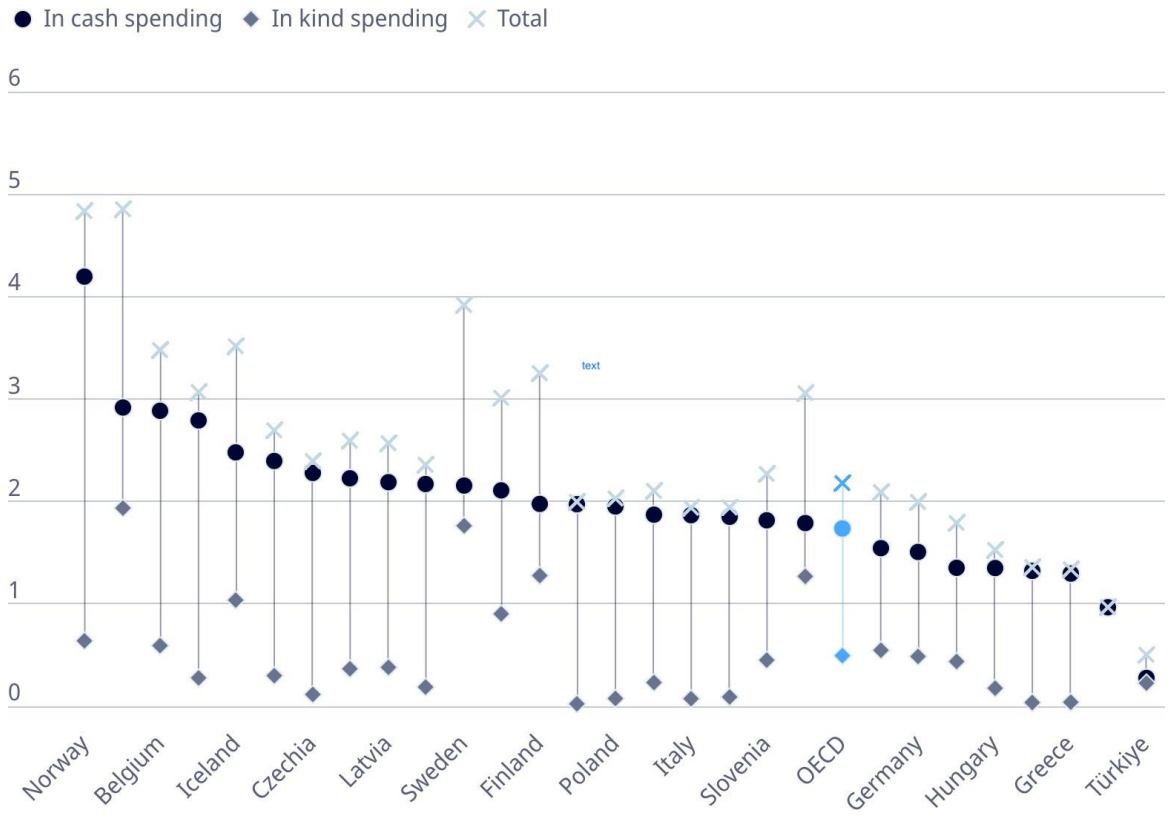


Panel B: Net entries on short-term DI



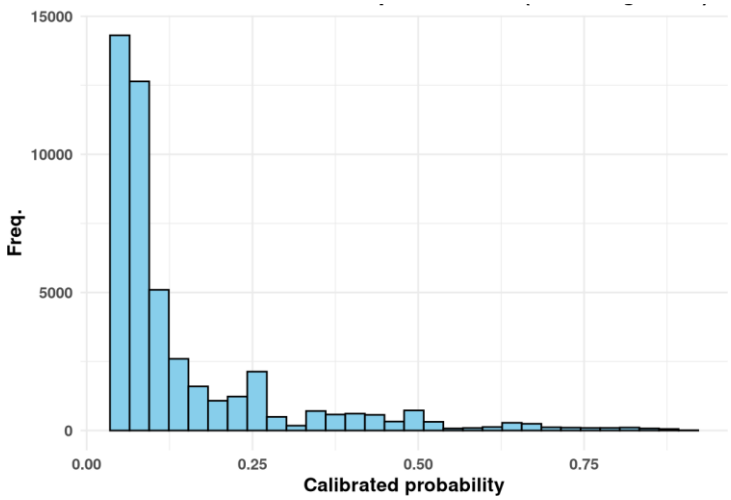
Note: self-constructed graphics with data from the National Institute for Health and Disability Insurance (NIHDI; INAMI in French). Panel A represent the total number of individuals on the long-term disability insurance program (more than 1 year) while panel B represents the net entries; the number of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

Figure A2: Public spending on incapacity in 2020



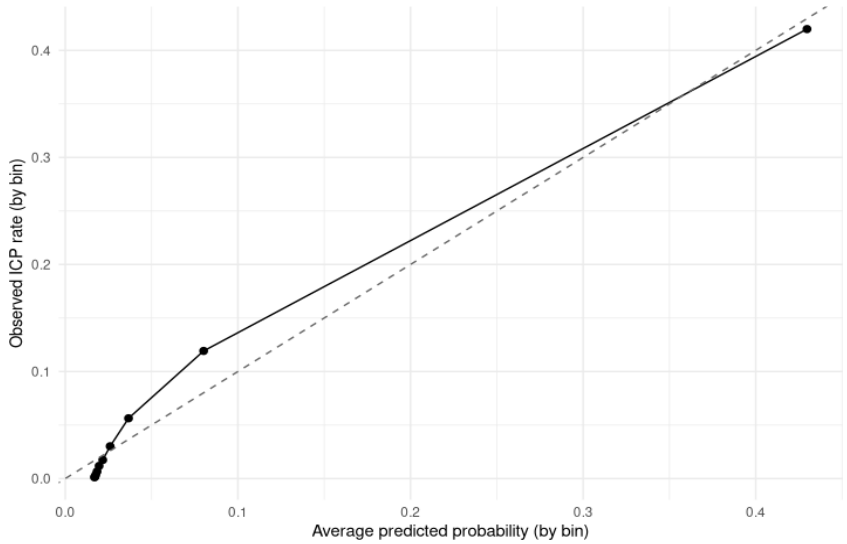
Note: Public spending on incapacity in % of GDP in the OCDE countries in 2020, it includes disability payments in cash, in kind and the sum of both. Data comes from OCDE statistics, 2020.

Figure A3: Distribution of predicted transition probability (excluding probabilities <0.05)



Note: The plot presents the distribution of the predicted work incapacity entry probability in the hold-out sample for the year 2018, excluding all the values lower than 0.05, to improve visual interpretation of the rest of the distribution.

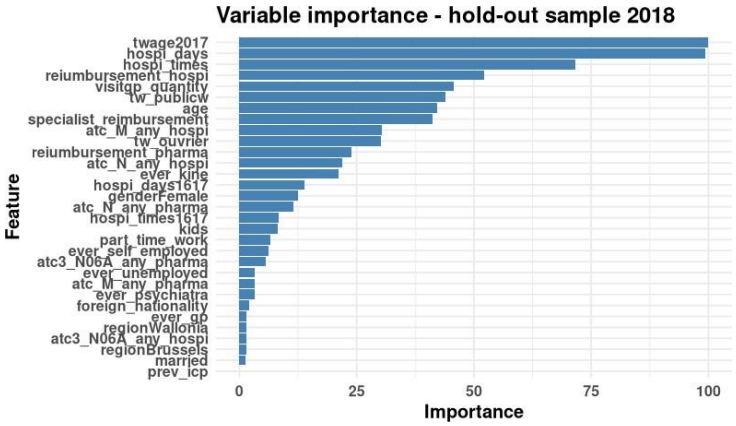
Figure A4: Comparing predictions from the baseline model to outcomes in 2018



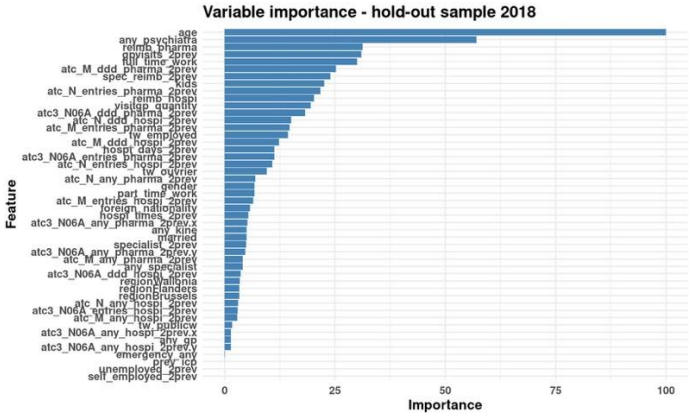
Note: Panel A shows a calibration plot comparing predicted and observed work incapacity entry in the hold-out sample for 2018, using all the individuals in our sample, without excluding the ones with previous work incapacity spells. Individuals are grouped into 10 deciles based on their predicted work incapacity entry probability. For each bin, we plot the average predicted probability against the observed work incapacity entry rate. The dashed line represents perfect calibration (i.e., predicted = observed). Because the plot uses decile averages, the values do not span the full 0–1 range even though individual predictions do.

Figure A5: Variables importance

Panel A: Work incapacity entry risk (excluding individuals with a previous DI spell)



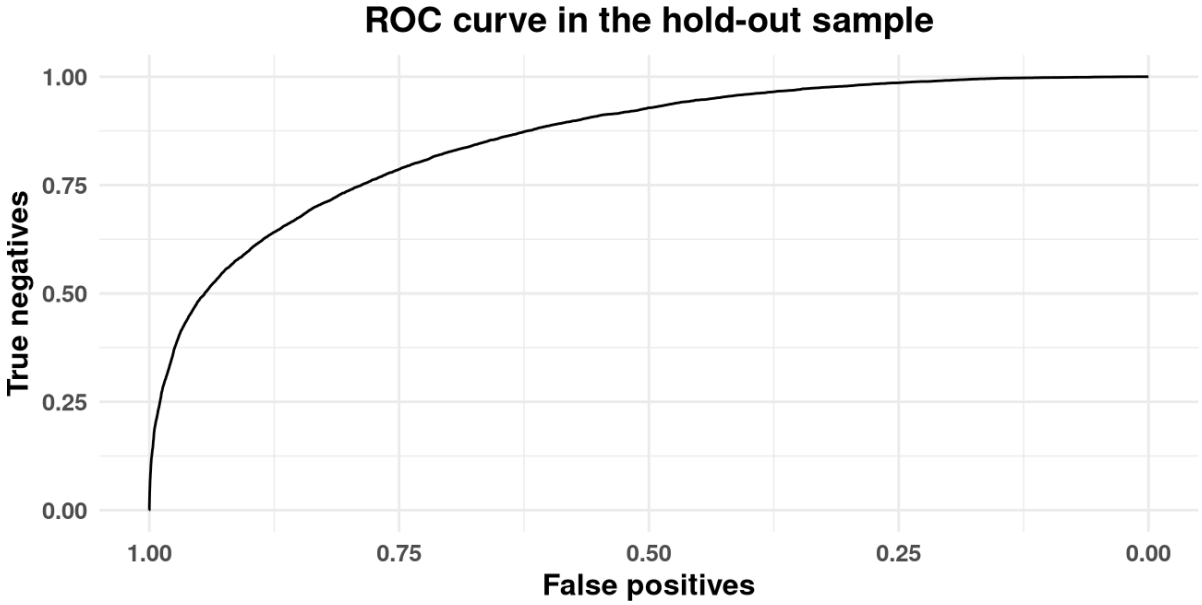
Panel B: Probability of transitioning from ST- to LT-DI



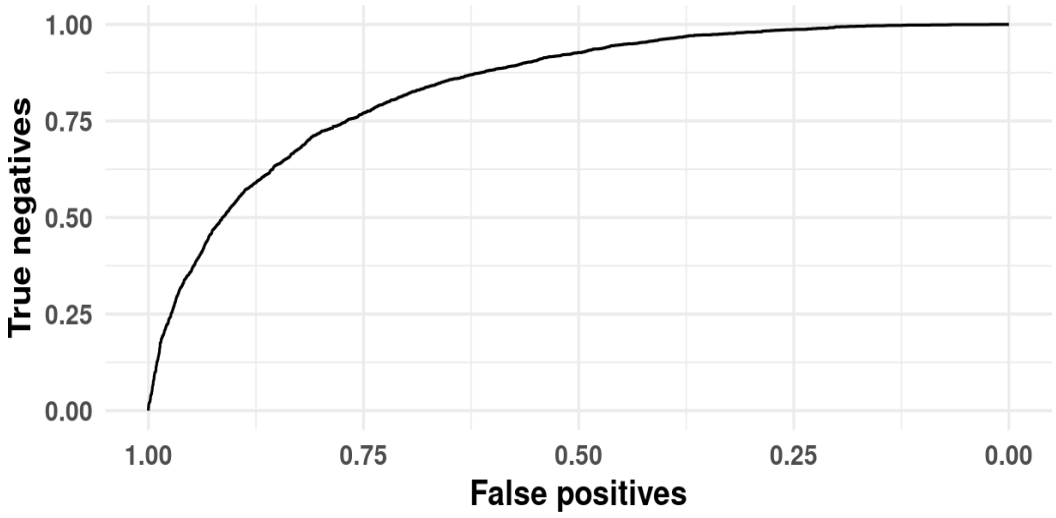
Note: Contribution of each variable to the predictive power of the three different ensemble models for 2018.

Figure A6: 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.

Table A1: AUC and R^2 for the models in Figure 6

Panel / Model	Specification	AUC	R^2
Panel A — Dynamic selection	0 → 2 quarters (baseline)	0.6410	0.0558
	2 → 4 quarters (conditional on staying)	0.6502	0.0855
Panel B — Forecasting horizons	1-quarter horizon	0.5765	0.0150
	2-quarter horizon	0.6410	0.0558
	3-quarter horizon	0.6565	0.0600
	4-quarter horizon	0.6265	0.0209
Panel C — Information sets	Reduced model (basic variables only)	0.5798	0.0209
	Full baseline model	0.6410	0.0558
Panel D — Cohorts	Cohort 2010	0.6524	0.0653
	Cohort 2018	0.6410	0.0558

Note: These metrics correspond to the predictive performance of the models illustrated in Figure 6. For each panel, the baseline case—representing the predicted two-quarter exit probabilities at the start of the spell—is reported again to facilitate comparison across specifications; this baseline distribution is also shaded in grey. The AUC quantifies the model's ability to discriminate between individuals who exit and those who remain in work incapacity, while the R^2 captures the proportion of variation in predicted exit risks explained by observable characteristics.

B. Data and Institutional 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. Our unit of observation is the individual–quarter, which allows us to construct quarterly spells of short-term DI and to observe transitions to long-term DI at the 12-month threshold.

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. In addition to modelling entry and transition probabilities, we later exploit the longitudinal nature of the dataset to track how individuals' status evolves within short-term DI spells. This allows us to study the probability of remaining in short-term DI after one, two, three, or four quarters, which forms the basis of the dynamic analysis. In this case, the outcome is defined as a dichotomous indicator equal to one if the individual remains in short-term DI for the corresponding number of consecutive quarters.

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.

C. 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.

C.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

$$E_t(\hat{R}_{i,t}|X_i) = D_t(X_i)$$

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.

C.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:

$$D_{i,t+1} = D_{t+1}(X_{i,t}) + \varepsilon_{i,t+1}, \qquad \mathrm{DY}\varepsilon_{i,t+1}|X_{i,t}\mathrm{Z} = 0$$

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.

C.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,

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

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

$$\bar{h}(d+1) - \bar{h}(d) = \mathrm{D}[\bar{h}_i(d+1) - \bar{h}(d) \,| i \in S_{d+1}] + (\,\mathrm{D}[h_i(d) | i \in S_{d+1}] - \mathrm{D}[h_i(d) | i \in S_d])$$

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.

Dynamic selection and duration dependence represent two distinct channels through which persistence in DI spells may arise. Dynamic selection reflects underlying heterogeneity: individuals differ in their exit hazards from the moment they enter DI, and those with higher hazards tend to leave early, progressively concentrating the remaining population among those with intrinsically lower recovery prospects. Duration dependence instead captures a causal effect of time spent in DI on subsequent exit probabilities, whereby remaining longer in the program reduces the likelihood of exit even for individuals with similar observables, potentially due to health deterioration, learning about the DI system, or increasing distance from employment. These mechanisms have opposite empirical implications: dynamic selection leads to narrower risk distributions and improved predictive accuracy as spells lengthen, while duration dependence generates patterns that observables cannot explain and therefore tends to reduce predictability.

Empirically, we distinguish duration dependence from dynamic selection by examining how predicted exit hazards and predictive accuracy evolve as DI spells progress. If duration dependence is present, individual exit hazards would decline with elapsed duration even after conditioning on observables, reflecting mechanisms such as clinical deterioration, psychological adaptation, or increasing detachment from the labor market. Such dynamics would make exit risks evolve in ways that the model cannot capture, thereby reducing predictive accuracy. In contrast, if dynamic selection dominates, individuals with higher exit hazards leave DI early, and those who remain over time become a more homogeneous group with systematically lower hazards. Under this mechanism, the distribution of predicted risks becomes more concentrated as spells lengthen, and predictive performance tends to improve

because observable characteristics become more informative for those who remain. To assess the relative importance of these two forces, we study how predicted exit risks and standard predictive metrics such as the AUC and R^2 behave at different elapsed durations and forecasting horizons.

D. Prediction Model: details on the 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 help 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. To further understand the dynamics for the first entry on DI, we exclude from the baseline sample individuals who had an ICP episode in the previous two years. 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 to different years 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 and the learning rate for the Boosted Regression Trees. 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:

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

where $\hat{p}_x$ is the prediction from algorithm x and βa_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 main results we present correspond to this hold-out evaluation for the year 2018. These prediction models are not designed for causal inference but to quantify the share of systematic variation in DI outcomes that can be captured by observables (i.e., the degree of observable heterogeneity).

Assessing the model:

To evaluate the accuracy of our prediction model, we compare predictions and outcomes in the hold-out sample for the year 2018 in Figure 1 of the main French report. 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 A of Figure 3 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.

We use our prediction model to assess the probability of entering DI and, conditional on entry, the probability of transitioning to long-term DI. Panel B of both figures display the distribution of calibrated predicted probabilities in the 2018 hold-out sample. In both panels, most of the predicted probabilities remain below 0.5. 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. Figure 1 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. 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. 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. Figure 3 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 A6 in Appendix A 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 0.86 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 0.189 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.

Dynamic prediction models:
Building on the baseline setup, we estimate a set of dynamic prediction models to study how predictive performance evolves as short-term work-incapacity spells unfold. Following the intuition in Mueller and Spinnewijn (2023), we examine how exit risks and their predictability change with elapsed duration and with different forecasting horizons. In this part of the analysis, the outcome of interest is not entry into DI but the probability of exiting short-term DI within a specified horizon.
We first estimate models at different durations of the spell (e.g., at the start of the spell and after two quarters) to assess whether predictive accuracy improves or deteriorates as individuals remain longer in DI. This allows us to infer the relative importance of dynamic selection versus duration dependence: increasing predictability with elapsed time is indicative of stronger dynamic selection, as individuals with higher exit hazards tend to leave early, leaving a more homogeneous population with respect to observed characteristics.
We then evaluate models with alternative forecasting horizons; predicting the probability of exit within one, two, three, or four quarters, while holding constant the information set available at the start of the horizon. Comparing predictive accuracy across horizons provides insight into how informative observables remain when forecasting persistence over different time frames.
All dynamic models use the same set of predictors as the baseline model, measured up to the relevant point in time, and are evaluated on independent hold-out samples. Together, these exercises characterise how the predictability of DI persistence evolves both with spell duration

and with the length of the prediction window, shedding light on the mechanisms shaping early patterns of return to work.

Predicting variables:

To understand how different groups of variables contribute to predictive performance, we estimate a sequence of nested models that progressively incorporate additional information. Prior work has shown that sociodemographic characteristics such as age or gender correlate with DI risks, yet these factors explain only a small share of the overall variation in disability recipiency (De Brouwer and Tojerow, 2023). This suggests that broader and more detailed sources of heterogeneity must be taken into account when analysing DI entry and persistence. A key strength of our setting is the richness of the administrative data, which allows us to combine sociodemographics with labour-market histories and multiple dimensions of healthcare utilisation. Table 1 summarises the variable groups used across the models, and Table 3 reports the predictive accuracy of the nested specifications. Consistent with the literature, sociodemographics alone provide limited predictive power, while adding labour-market information leads to a substantial improvement. The largest gains arise when health-related variables are included, underscoring their central role in structuring work-incapacity risks.

In the machine-learning models, we use variable-importance measures to assess the marginal contribution of predictors to model performance. These measures, shown in Figure A5 of the Appendix, capture both non-linearities and interaction effects, which are often substantial for health-related variables (e.g. hospitalisations or pharmaceutical expenditures). However, they do not have a causal interpretation, nor do they provide easily interpretable linear associations. For this reason, and following the approach of Mueller and Spinnewijn (2023), we complement the ML variable-importance results with a standardized OLS regression of the calibrated predicted probabilities on the underlying predictors. This regression, reported in Figures 2 and 4 for the first two outcomes, expresses coefficients in standard-deviation units and offers a transparent measure of how each observable characteristic is linearly associated with the predicted risk. Unlike the ML importance metrics, the OLS coefficients summarise only linear relationships and do not reflect interactions or non-linear effects. As a result, the two approaches naturally yield different rankings of predictors, but together they provide a comprehensive and interpretable picture of the observable factors shaping work-incapacity risks.