Title

When (and What) Matters in the Late-Career Path? A Diagnostic Sequence Approach Using Transformers to Locate Poverty Risk in Late-Career Histories

Linda Vecgaile (Max Planck Institute for Demographic Research)

Work in Progress

Keywords: poverty risk; late career; ageing and work; Germany; United Kingdom; sequence methods; transformer; explainability; early warning

Abstract

European welfare states increasingly prioritise extended working lives and recognise the need to support individuals in the late-career phase - often from around age 55 - to prevent material hardship as workers approach pension age. A practical design question remains open: when and where to intervene to reduce near-term poverty risk. We study a late-career support horizon (ages 55–59) and its association with poverty at ages 60–62 - a pre-retirement period in which shocks can compound and shape income trajectories into retirement.

We treat a Transformer model as a measurement device to locate where predictive signal resides within late-career histories. Using CPF-harmonised panel data from Germany and the United Kingdom, we encode yearly labour-market states (employment, hours, sector, industry, occupation, disability) and predict poverty at ages 60–62. We then assess when in the 55–59 window information is most informative for the predicted poverty and which features matter most. In addition, we run calibrated "what-if" edits to estimate how shifting disability to no disability would change predicted risk. Finally, we document how these patterns differ by country and by gender.

Three results stand out. Among those who become poor, late-window disability is the principal predictor - especially in the United Kingdom and for women. Among those who avoid poverty, stable working time is the dominant protective margin. Across diagnostics, signal concentrates at ages 58–59. Consistent with this, flipping disability from "yes" to "no" lowers predicted risk in the UK. We make no causal claims; rather, we offer a transparent diagnostic framework indicating when support is most informative (58–59) and what to prioritise (late disability risk vs. working-time stability), to complement causal evaluations in European social policy.

1. Introduction

Across Europe, pension and labour-market reforms have encouraged later retirement and longer working lives. Yet a non-trivial share of workers struggle to maintain stable employment in late-career years due to skill obsolescence, health limitations, discrimination, or repeated job-loss and hours shocks. When setbacks arrive late in the career, they are harder to reverse. As a result, older

displaced workers suffer persistent earnings losses and lower re-employment rates than mid-career peers (Jacobson, LaLonde and Sullivan, 1993; Chan and Huff Stevens, 2001). Comparative reviews emphasise that recovery chances fall sharply after the mid-50s and that late-career shocks can cause exits that bridge into retirement, raising poverty risk (OECD, 2010, 2019).

European welfare states increasingly target support to the late-career phase (often from around age 55), but policy design still lacks precise guidance on when and where to intervene and for how long. Classic models identify correlates of poverty risk but do not localise when in a short late-career sequence the information resides, nor whether a one-year push, two–three years, or sustained support across the window is more informative.

We bring a diagnostic sequence perspective to this problem. A Transformer is used as a measurement device to locate where, in late-career histories, predictive signal lives. Using CPF-harmonised panel data (Turek, Voets and Kalmijn, 2023) for Germany and the United Kingdom - countries with contrasting pension architectures (Börsch-Supan and Ludwig, 2010; Cribb and Emmerson, 2020; Scarpetta and di Noia, 2023; Murphy, 2024) - we focus on ages 55–59 and predict poverty at 60–62. Conditions in these final working years shape exposure to income loss and the resources carried into retirement via earnings histories, contribution records, and eligibility thresholds. Reducing poverty risk in this period can help stabilise trajectories into statutory pension ages.

We make no causal claims. Instead, we offer a transparent measurement framework that narrows where to test causally, complements programme evaluations, and supports early-warning and targeting as welfare states operationalise extended working lives.

2. Literature review

Research on ageing and work consistently shows that shocks late in the career, such as job loss, sustained hours reductions, and the onset or worsening of health limitations, are unusually hard to reverse and carry long-lasting financial consequences. Older displaced workers experience large and persistent earnings losses and markedly lower re-employment probabilities than mid-career peers (Jacobson, LaLonde and Sullivan, 1993; Chan and Huff Stevens, 2001). Comparative policy reviews for Europe emphasise that the return-to-work chances fall steeply after the mid-50s and that late-career setbacks can result in exits that bridge into retirement or inactivity and raise poverty risk (OECD, 2010, 2019). Disability and poor health amplify these dynamics by reducing employment probabilities and hours at older ages (OECD, 2010; Börsch-Supan, 2013).

Institutional context conditions how these late-career events translate into near-retirement income security. Germany's public pension is earnings-related and built on contribution points, so covered employment and relative earnings (especially the density and level recorded in later years) directly shape pension entitlements (Börsch-Supan and Ludwig, 2010; Scarpetta and di Noia, 2023). The United Kingdom's New State Pension is closer to flat-rate, with late-career labour-market conditions affecting short-run income and private accumulation more than the state benefit itself. Since automatic enrolment (post-2012), defined-contribution workplace pensions have become the dominant channel for late-career saving. Therefore, sustained work and sufficient hours in the late 50s matter for contribution flows and the liquidity households carry to state pension age (Cribb

and Emmerson, 2020; Murphy, 2024). In both settings, late-career shocks can deplete buffers before eligibility for full old-age benefits, and the scope for recovery narrows sharply in the early 60s.

A second strand of literature pinpoints the specific ingredients of late-career income stability. Employment continuity is the most robust protective factor: older workers who sustain attachment avoid the steep scarring seen after displacement and accumulate either contribution points (Germany) or pension savings and liquidity (UK) (Chan and Huff Stevens, 2001; Cribb, Hood and Joyce, 2017). Hours matter as well - not only mechanically for earnings but also because part-time or irregular schedules are associated with weaker attachment trajectories at older ages (OECD, 2019). Sectoral location contributes through job security and displacement risk: public-sector employment typically offers lower dismissal risk, more stable hours, and stronger protection against late-career scarring than otherwise similar private-sector jobs as documented in European labour-market studies (Boeri, 2010). Finally, disability and poor health are consistently linked to later poverty via reduced employment probabilities, restricted hours, and earlier labour-market exit (OECD, 2010; Börsch-Supan, 2013).

While this literature identifies what correlates with near-retirement poverty risk, it is less informative about when in a short late-career sequence the relevant information resides, and whether a brief late push or multi-year support is more informative. That timing question has become central as European welfare states seek to operationalise "extended working lives" policies. Recent work in prediction-for-policy argues for tools that can support targeting and early warning without over-claiming causality (Kleinberg *et al.*, 2015). In parallel, explainable machine-learning methods, especially perturbation-based approaches, offer ways to interpret complex models. Global importance scores (e.g., SHAP: (Lundberg and Lee, 2017)) summarise average associations but do not localise signal to specific years within short sequences. Domain-aligned interpretations such as sliding window masks, leave-one-age-out ablations, and calibrated token edits provide a transparent way to map where model-detected information resides along the path and which feature families supply it. Used as a diagnostic rather than a causal tool, this "sequence-aware" perspective helps identify the ages and features most consequential for near-term poverty risk and can narrow the design space for subsequent quasi-experimental or experimental evaluations.

3. Data and outcome

We use the CPF harmonised longitudinal file for Germany and the United Kingdom (Turek, Voets and Kalmijn, 2023). The CPF project standardises core labour-market, income and demographic information across panel surveys and releases a person-period file with consistent labels and value coding. From this source we construct short late-career sequences and a near-term poverty outcome.

The analytic population consists of individuals who are observed continuously from ages 55 to 59 with sufficient information to form annual labour-market tokens (defined below), and for whom poverty status can be determined at ages 60–62 that comprises over 1300 individuals. The unit of analysis is the person; sequences are built over five late-career ages (55–59) and the outcome is evaluated in the subsequent three-year window (60–62). All diagnostics reported in the paper use

the validation split for transparency and reproducibility; final summary figures are replicated on a held-out test split in the Supplement.

The outcome is a binary indicator equal to one if the respondent is recorded as poor in any of the ages 60, 61 or 62 ("any poverty 60–62"), and zero otherwise. In the CPF harmonisation, poverty is derived from equivalised post-tax household income using the standard country-year poverty line adopted in the underlying survey (consistent with LIS/Eurostat conventions); we follow that harmonised definition and carry it through unchanged. This target is substantively motivated: 60–62 is a high-stakes pre-retirement interval in which shocks are difficult to undo and can shape resources carried into statutory pension ages. Choosing a three-year window avoids classifying transient monthly fluctuations as poverty while preserving sensitivity to sustained short-run hardship. Exact operational details (income measure, equivalence scale and threshold used in each source survey) are documented in Supplement A.

For each age $a \in \{55,56,57,58,59\}$ we emit a set of categorical tokens that summarise the person's labour-market state. Tokens are grouped into seven "feature families":

- E — Employment status (employed, unemployed/active, retired/disabled, not active/home, in education), derived from *emplst5/emplst6* and *work_d*.
- H — Working-time status (full-time vs part-time/irregular), from *fptime_h* / contracted or reported hours.
- D — Disability (any vs none), from *disab* (with *disab2c* used as a robustness check).
- P — Employer sector (public vs private), from *public*.
- I — Industry (one-digit major groups), from *indust1* (with *indust2/3* used only in sensitivity checks).
- O — Occupation (one-digit ISCO), from *isco_1* (harmonised across ISCO-88/08).
- W — Weekly hours (binned), from *whweek* (or *whweek_ctr* when only contracted hours are available).

These families are the ones we analyse throughout (age×feature ablations, class-conditional ΔBrier, and token-edit simulations). They were selected ex ante on three grounds that matter for policy design and for comparability across the two countries: (i) actionability (each maps to a lever or constraint that late-career supports can plausibly target such as employment attachment, hours accommodation, sectoral placement, disability mitigation, or reallocation across broad industries/occupations); (ii) coverage and harmonisation (all seven are present and consistently coded in both Germany and the UK in the CPF); and (iii) parsimony (they provide broad behavioural and institutional characteristics without overfitting to country-specific micro-categories). When a token family is not observed at a given age (e.g., missing sector among non-workers), a "–MISS" variant is recorded so that the model does not impute content that is not present.

Country and gender appear as header tokens (not part of the age-stamped sequence) using the harmonised *country* and *female* fields. They enable us to stratify every diagnostic by country×gender without leaking outcome information into the sequence itself.

Why these predictors (and not others)?

The CPF file includes a rich set of additional variables (education, detailed income components, self-rated health, satisfaction scales, family composition, parental background, migration, religiosity, etc.; a full inventory is listed in Supplement B). We deliberately do not include them in the main sequence for three reasons.

1. Temporal fit to a short late-career horizon. Several variables are time-invariant or slow-moving at ages 55–59 (e.g., highest education, parental background). Including them inside a five-year sequence would blur the timing logic that our diagnostics are designed to recover. Where substantively important (e.g., education), we report sensitivity analyses that add them as baseline controls rather than age-varying tokens (Supplement C).
2. Comparability and coverage. Some candidate predictors (e.g., detailed firm size, fine industry/occupation, satisfaction scales) exhibit differential availability or harmonisation across Germany and the UK. To avoid artefacts driven by country-specific missingness patterns, we restrict the main analysis to families with stable cross-country coverage. We document coverage and missingness by feature family in Supplement B.
3. Avoiding post-outcome contamination. Detailed contemporaneous income variables and benefit receipts are close to the poverty construct itself. Using them as sequence tokens risks tautology. Our labour-market tokens capture *conditions* plausibly upstream of poverty without leaking the outcome into the inputs.

This design choice is conservative: it prioritises interpretability of when and what in short late-career histories over maximal predictive accuracy from a high-dimensional feature set. In return, it lets us read the trained model as a measurement device that localises policy-relevant information in a way that is comparable between Germany and the UK.

Because our objective is to diagnose where signal resides rather than to estimate population levels, model training and perturbation diagnostics are conducted without survey weights; descriptive tables of the analytic sample are provided with recommended survey weights in Supplement D. The "val" split is used for all figures in the main text to keep the workflow reproducible; analogous figures for the held-out test split appear in Supplement E.

3. Methods: the Transformer as a diagnostic measurement device

We use a sequence Transformer to predict poverty at ages 60–62 from annual tokens observed at 55–59. Following work that repurposes complex learners as measurement devices (not just prediction engines) we read the trained model to localise where the predictive signal resides in short life-course segments (Breiman, 2001; Doshi-Velez and Kim, 2017; Rudin, 2019). The aim is interpretability of timing and content, not structural causality.

3.1 Why a Transformer fits life-course data

The Transformer is a natural fit for life-course questions because the model encodes ordering and recency through positional encodings and self-attention, letting the model weight events at different ages relative to each other (Vaswani *et al.*, 2017). This maps onto classic life-course principles: outcomes are shaped by when events happen, how long they persist, and in what sequence they occur (Abbott, 1995; Elder, 1998; Dannefer, 2003; Kuh *et al.*, 2003).

Related ML work has used mask-based perturbations and window searches to localise influential time spans (Tonekaboni *et al.*, 2019; Ismail, Corrada Bravo and Feizi, 2021), feature-time ablations in clinical sequences (Suresh, 2017), and sequence-level counterfactual edits that change spans and examine prediction movement (Ross, Hughes and Doshi-Velez, 2017; Goyal *et al.*, 2019; Delaney, Greene and Keane, 2021). Our set borrows these ideas but tailors them to a short, policy-relevant horizon (55–59) and to outcomes central to European social policy. Our goal is interpretability of timing and content, not structural causality.

3.2 Tokenisation

For each person and age $a \in \{55,56,57,58,59\}$, we emit tokens from six labour-market families: employment status (E), working-time status (H), disability (D), sector public/private (P), one-digit industry (I), and one-digit ISCO occupation (O), plus weekly-hours bins (W). Country and gender are header tokens. The target is any household poverty at ages 60–62 (binary). Sequences are fed to a text-vectorisation pipeline and an encoder–decoder Transformer classifier (details in Supplement A).

3.3 A coordinated diagnostic set

We treat the trained model as an instrument that we probe by perturbation. Each diagnostic asks a concrete "where/what/for whom?" question and reads off the model's response using calibration and positive-class metrics appropriate for imbalanced screening (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015).

(i) Sliding-window masks: where is the information?

We slide one-year and two–three-year masks across ages 55–59 and recompute Brier score; the loss in fit (ΔBrier = masked − full) indicates when the sequence carries the most information.

(ii) Age×feature ablations: what matters when?

For each feature family × age, we mask tokens and recompute ΔBrier. Larger ΔBrier implies the model relied on that feature×age for calibration. We show panels for ALL, and by country×gender.

(iii) Positive-class diagnostics: who do we catch?

Because policy screening focuses on identifying future poor, we compute for each feature×age:

- ΔPR-AUC (change in average precision for the positive class), and
- ΔTPR at calibrated thresholds (per-country; ≈0.30 DE, ≈0.23 UK/ALL from prior threshold sweeps).
  These metrics isolate which feature×age helps catch the positive class.

(iv) Class-conditional ΔBrier: risk-raising vs risk-lowering content

We recompute ΔBrier within positives (y=1) and within negatives (y=0). This disentangles features that maintain high risk among those who will become poor (e.g., late disability) from features that keep risk low among those who remain non-poor (e.g., sustained employment).

(v) Calibrated token-edit simulations: directional "what-if"

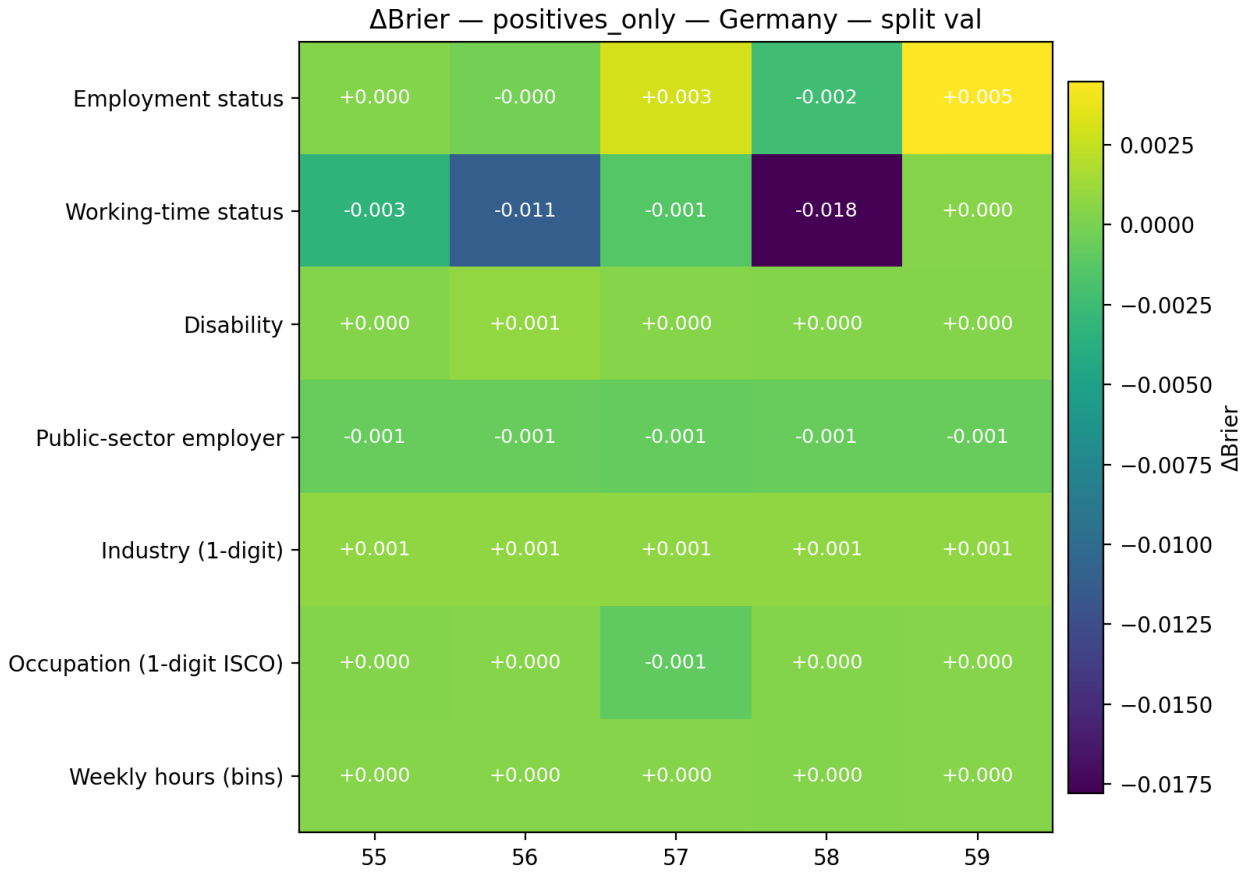
We change tokens in place over pre-specified windows and re-score risk:

- Disability yes to no (main text), and
- Sector flips private to public / public to private (supplement).
  For windows 58–59, 55–57, and 55–59, we report, by group:
  (i) mean Δrisk (edited − original) with bootstrap 95% CIs, and
  (ii) % crossing the screening threshold from below (proxy for newly surfaced borderline cases).

4. Results

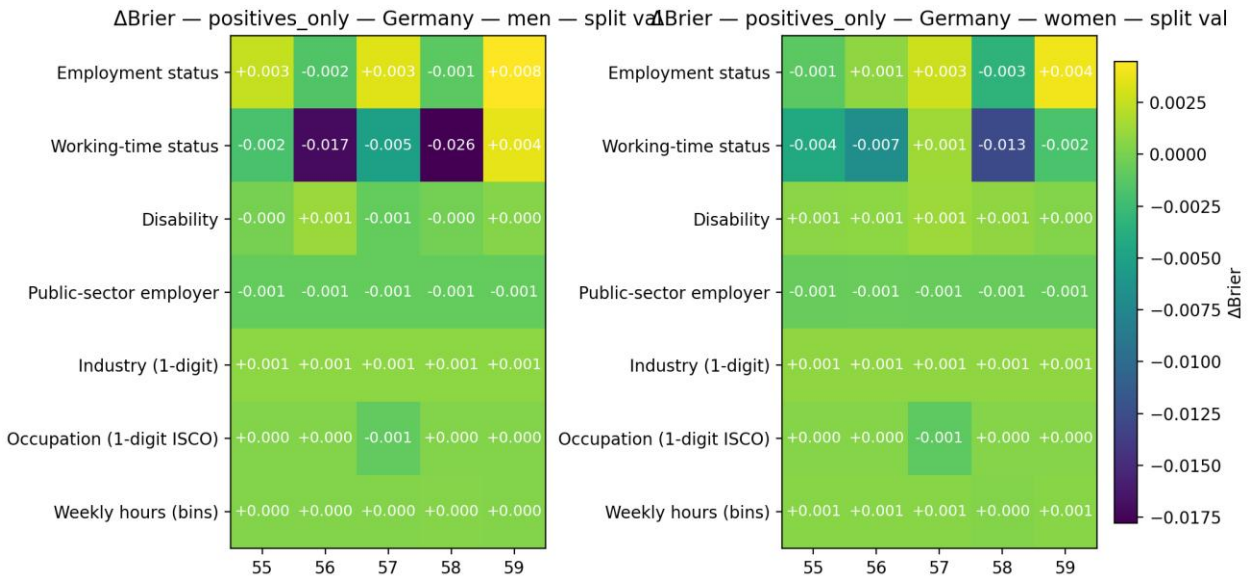
We begin by restricting the evaluation set to respondents who subsequently become poor. In the United Kingdom (Fig. 1-2), disability carries a clear late-window signal (ΔBrier ≈ +0.004 at ages 57 and 59), with the largest effect among women (≈ +0.005 at age 59). In Germany (Fig. 3-4), disability remains informative but more modest (≈ +0.001 around ages 56–57). By contrast, working-time status does not help identify which members of the positive class cross into poverty; its ablation often improves calibration, especially at age 58 (ΔBrier ≈ −0.018 in Germany; ≈ −0.012 in the UK), indicating that hours information is not the separating margin in this subset. Instead, for Germany a late employment signal emerges: removing employment status at age 59 worsens calibration (ΔBrier ≈ +0.005 overall; ≈ +0.008 for men), suggesting that continued attachment at the very end of working years helps keep predicted risk high among those who will in fact experience poverty. Sector, industry, occupation, and weekly-hours bins are consistently small in magnitude throughout these positives-only panels, reinforcing the centrality of late-window disability, and, in Germany, end-of-window employment status -for flagging those at greatest risk.

Figure 1. ΔBrier (positives only) - Germany, ages 55–59



Heatmap of change in Brier score (ΔBrier) when a single feature is masked at each age, restricting the evaluation set to respondents who later become poor (positives only). Rows are feature families; columns are ages 55–59. Positive values (warmer colours) mean calibration worsens when the feature is removed → the feature carries informative signal at that age. Negative values (cooler colours) mean masking improves calibration. In Germany, disability contributes modestly (≈ +0.001 around 56–57), working-time often shows negative effects at 58 (≈ −0.018), and employment at 59 is informative (≈ +0.005).

Figure 2. ΔBrier (positives only) - Germany by gender

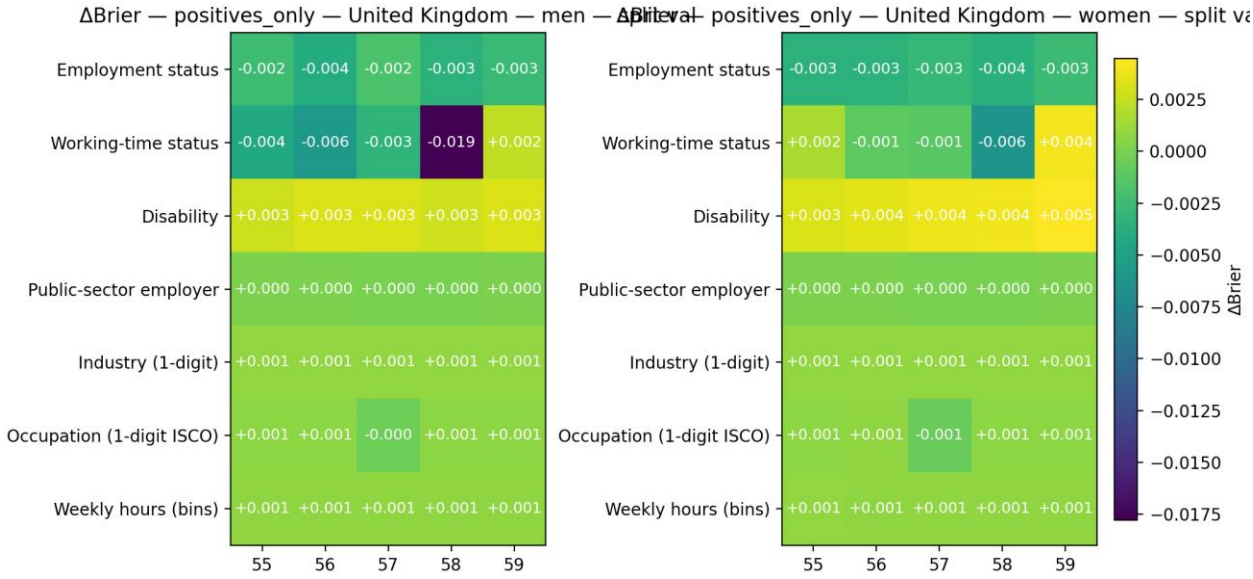


Heatmap of change in Brier score (ΔBrier) when a single feature is masked at each age, restricting the evaluation set to respondents who later become poor (positives only). Rows are feature families; columns are ages 55–59. Positive values (warmer colours) mean calibration worsens when the feature is removed → the feature carries informative signal at that age. Negative values (cooler colours) mean masking improves calibration. Among German men, late employment at 59 is most informative (≈ +0.008) and working-time at 58 is strongly negative (≈ −0.026). Among German women, patterns are similar but smaller in magnitude (e.g., working-time ≈ −0.013 at 58). Positive ΔBrier denotes loss of calibration when ablated; negative denotes improvement.

Figure 3. ΔBrier (positives only) - United Kingdom, ages 55–59

Heatmap of ΔBrier under age-specific feature ablations for UK respondents who later become poor. Late-window disability is the dominant signal (≈ +0.004 at ages 57 and 59). Working-time often shows negative effects, especially at 58 (≈ −0.012). Interpretation is as above: higher ΔBrier implies the masked feature was informative for calibration at that age.

Figure 4. ΔBrier (positives only) - United Kingdom by gender



UK positives-only ablations by sex. Disability is particularly informative for women in the late window (≈ +0.005 at 59), while men show a pronounced negative working-time effect at 58 (≈ −0.019). Rows are features; columns are ages 55–59; positive ΔBrier = worse calibration when masked (feature is informative); negative ΔBrier = improved calibration.

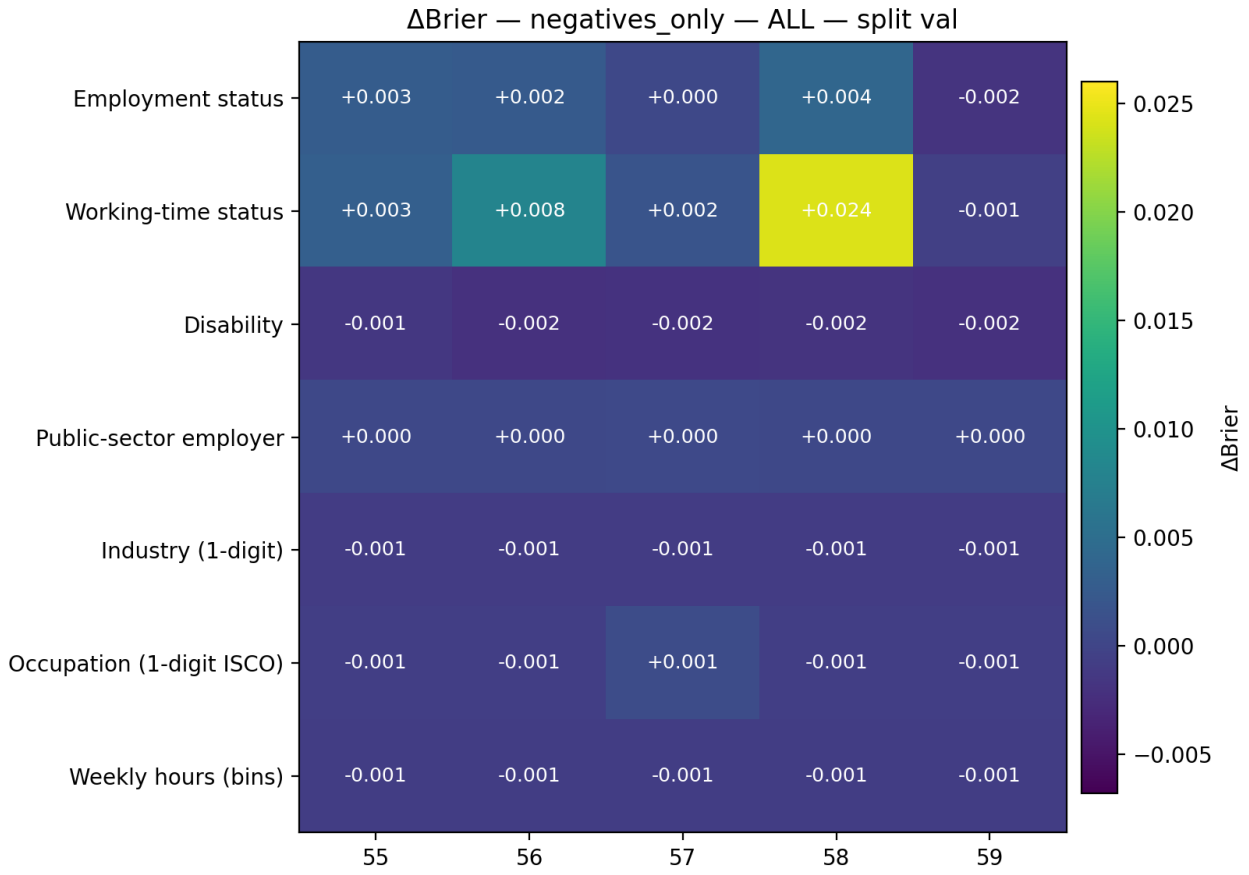
Our positives-only diagnostics align with life-course theory and the institutional contrast between Germany and the UK. Late health shocks are theorised to be hard to reverse; accordingly, ablation of disability at ages 58–59 produces the sharpest calibration loss among those who become poor (strongest in the UK, especially for women), indicating that new or worsening limitations near exit are pivotal for poverty transitions.

Country patterns reflect system design. In the United Kingdom, where near-retirement living standards depend on current earnings, liquidity, and DC contributions before State Pension Age, late-window disability is the dominant separator of the future poor. In Germany, with earnings-related pensions and stronger insurance tied to covered employment, the decisive separator within the positives is end-of-window attachment (notably at age 59), particularly for men in tenure- and strength-intensive jobs. Hours play a secondary role among the positives as many are already on reduced or unstable hours, so the transition is triggered less by marginal changes in working time than by a late disability shock or a final break in employment.

Further, we isolate the “negatives”, individuals who do not become poor at 60–62, and ask which late-career tokens keep predicted risk low. The pattern differs from the positives. In the pooled sample (Fig. 5), the dominant protective signal is working-time status, with a pronounced late-window peak: masking the working-time token at age 58 yields the largest deterioration in calibration (ΔBrier ≈ +0.024), with smaller but still visible losses at 56–57. Employment status contributes, but more modestly and without the sharp late-window concentration. By contrast, disability and sector carry little stabilising information among the negatives - their ablations hardly move calibration - suggesting that, for households who avoid poverty, maintaining stable hours

and attachment matters more than the presence or absence of disability records or public-/private-sector labels in this short horizon.

Figure 5. ΔBrier (leave-one-out) by feature×age - negatives-only, pooled (val split)



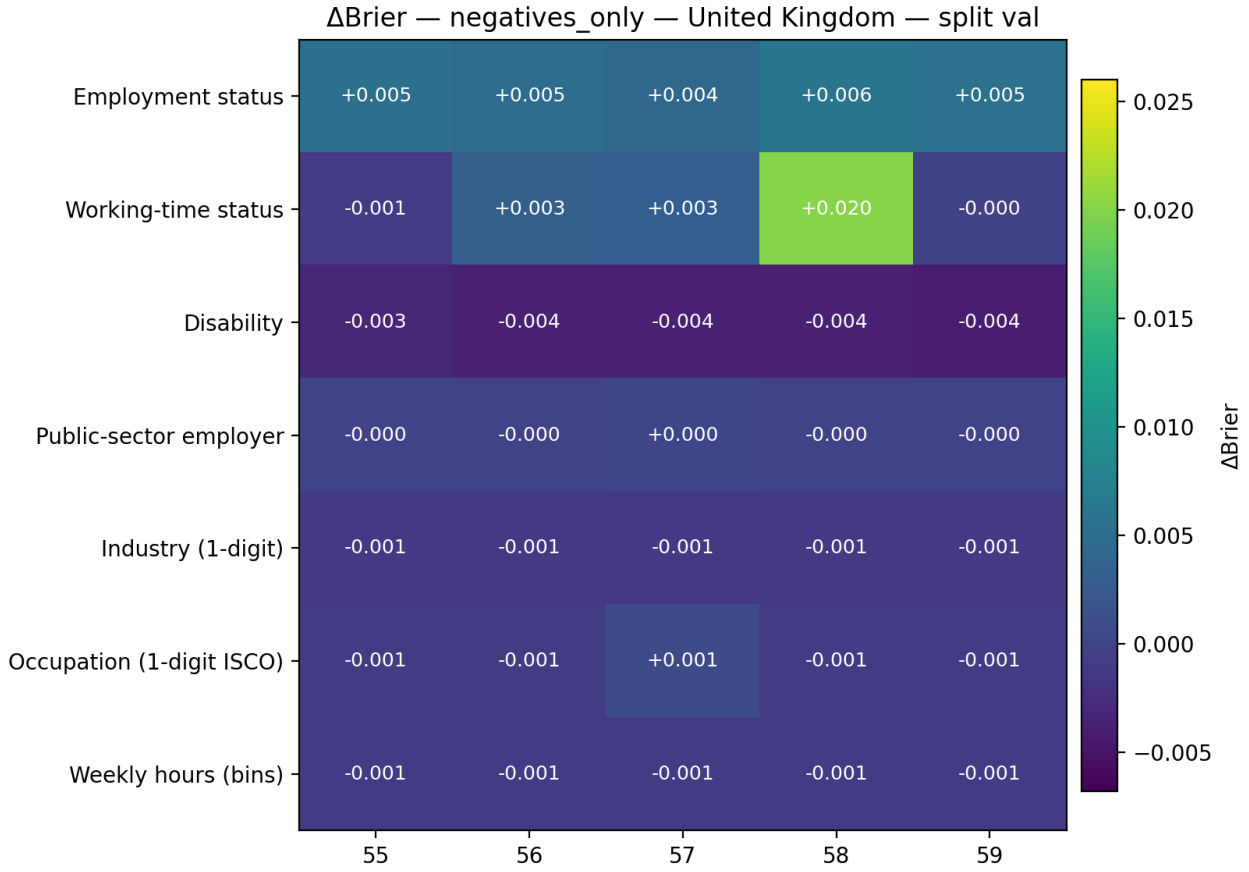
Heatmap shows the change in Brier score (ΔBrier) when the indicated feature at a given age (columns 55–59) is masked for the negative class (people who do not become poor). Cell values (and colours) are mean ΔBrier: positive = worse calibration after masking → the feature helps keep predicted risk low among negatives; negative = better calibration after masking → the feature adds little or hurts calibration. Working-time status at age 58 produces the largest loss (ΔBrier ≈ +0.024), indicating it is the dominant protective signal in the pooled data.

Country profiles reinforce this message. In Germany (Fig. 6), the hours signal is especially strong and late-loaded: removing working-time at age 58 produces the largest calibration loss (ΔBrier ≈ +0.027), with clear secondary peaks at 55–56. This aligns with the earnings-related architecture, where sustained, covered work-and in practice, stable full-time schedules-preserve contribution density just before pension claiming. In the United Kingdom (Fig. 7), hours again dominate with a late spike, consistent with the role of current earnings and hours for liquidity and DC contributions before State Pension Age. In both countries, ablations of sector and industry remain comparatively inert once hours and attachment are held fixed.

Figure 6. ΔBrier by feature×age — negatives-only, **Germany** (val split)

Heatmap shows the change in Brier score (ΔBrier) when the indicated feature at a given age (columns 55–59) is masked for the negative class (people who do not become poor). Cell values (and colours) are mean ΔBrier: positive = worse calibration after masking → the feature helps keep predicted risk low among negatives; negative = better calibration after masking → the feature adds little or hurts calibration. Larger positive ΔBrier marks features that most help the model correctly keep non-poor cases at low risk. The strongest effect is working-time at age 58 (ΔBrier ≈ +0.027), with smaller positive contributions at 55–56. Disability and sector rows are near zero, implying limited stabilising value once hours and attachment are held constant.

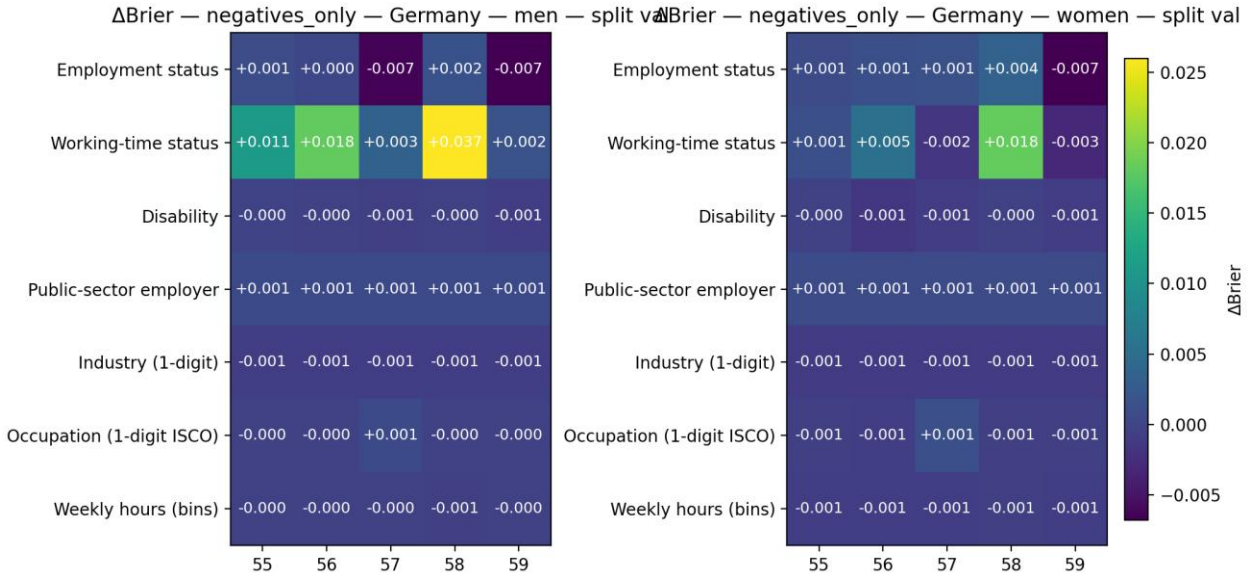
Figure 7. ΔBrier by feature×age - negatives-only, **United Kingdom** (val split)



Heatmap shows the change in Brier score (ΔBrier) when the indicated feature at a given age (columns 55–59) is masked for the negative class (people who do not become poor). Cell values (and colours) are mean ΔBrier: positive = worse calibration after masking → the feature helps keep predicted risk low among negatives; negative = better calibration after masking → the feature adds little or hurts calibration. Larger positive ΔBrier marks features that most help the model correctly keep non-poor cases at low risk. As in Germany, working-time at age 58 dominates (ΔBrier ≈ +0.020), while disability and sector rows remain close to zero or slightly negative. Employment status adds a modest, diffuse signal across ages.

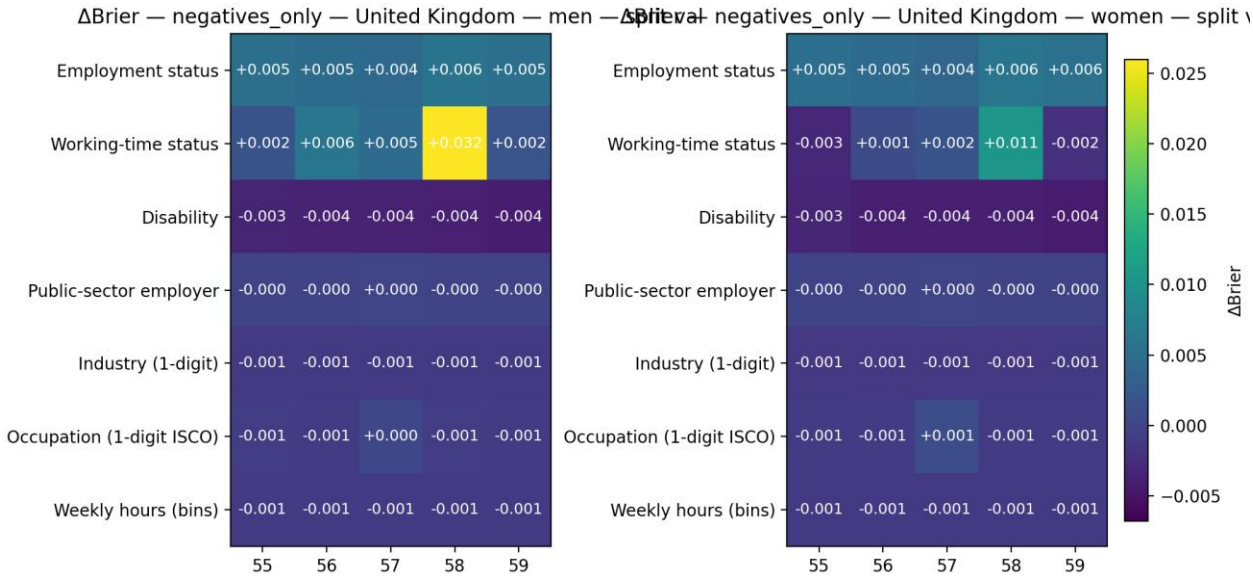
Gender splits amplify these cross-country regularities. Among German men (Fig. 8), working-time is the clearest protective margin, with a very large late-window peak (ΔBrier ≈ +0.037 at 58) and notable signal already at 55–56; employment status adds little by comparison. German women show the same ordering but with smaller magnitudes (ΔBrier ≈ +0.018 at 58), reflecting both higher part-time prevalence and the buffering role of stable schedules rather than sector per se. In the UK (Fig. 9), the late-window hours peak is again strongest for men (ΔBrier ≈ +0.032 at 58), while women exhibit a smaller hours peak (≈ +0.011) alongside a slightly larger contribution from steady employment at 58–59.

Figure 8. ΔBrier by feature×age - negatives-only, **Germany, men vs. women** (val split)



Panels repeat the German analysis by gender. For men, the working-time peak at 58 is very pronounced (ΔBrier ≈ +0.037); employment status contributes little by comparison. For women, the ordering is similar but magnitudes are smaller (working-time ≈ +0.018 at 58). Interpretation of signs matches prior figures: positive values indicate features that keep predicted risk low within the negatives.

Figure 9. ΔBrier by feature×age - negatives-only, **United Kingdom, men vs. women** (val split)



Gender split for the UK. Among men, working-time at 58 shows the largest protective effect (ΔBrier ≈ +0.032); women exhibit a smaller hours peak (≈ +0.011 at 58) alongside a slightly larger contribution from employment status at 58–59. Reading is as before: higher (more positive) ΔBrier = greater importance for keeping non-poor predictions well-calibrated.

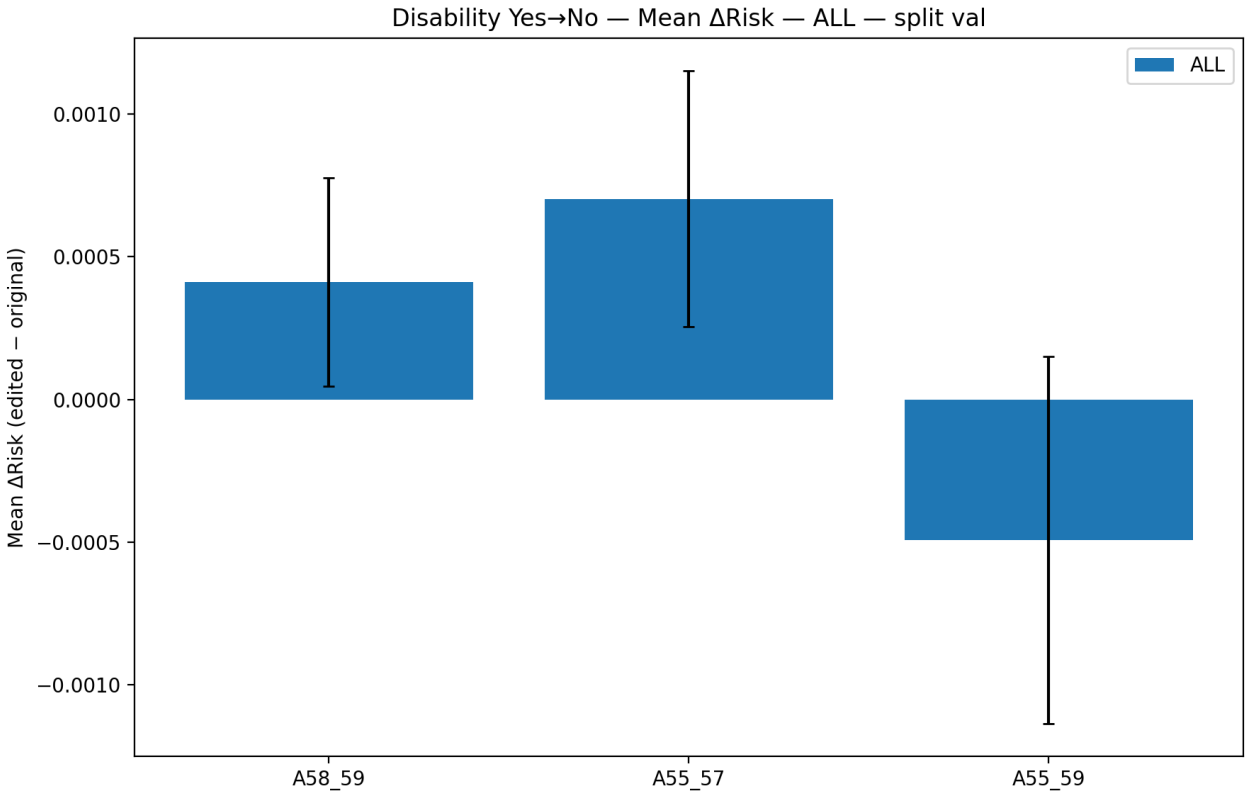
Taken together, these negatives-only diagnostics are tightly aligned with theory and institutions: households that avoid poverty do so by sustaining hours and attachment across 55–59, particularly in the last two years. In Germany this reads as preserving earnings-related accruals; in the UK it supports ongoing earnings and DC flows that maintain liquidity to SPA. The weak sector/disability signal among negatives is also expected: once hours and attachment are stable, sector labels or the absence of disability flags add comparatively little to near-term protection.

People who sustain full-time hours late in the 50s tend to be healthier, more experienced, and (on average) more educated. "Working-time status" is therefore a bundle: it captures hours and correlated advantages (job quality, tenure protection, union coverage, firm size, etc.).

4.2. Token-edit simulations (disability yes→no)

Counterfactual edits that switch disability from "yes" to "no" produce country-specific risk movements that align with our class-conditional diagnostics but also reveal composition and institutional differences. Pooled across countries (Fig.10), the edit lowers predicted poverty risk on average when applied to the full late-career window (A55–59), with smaller and imprecise effects for A58–59 and a modest increase for A55–57.

Figure 10. Counterfactual edit (Disability yes→no): mean change in predicted poverty risk - **ALL**

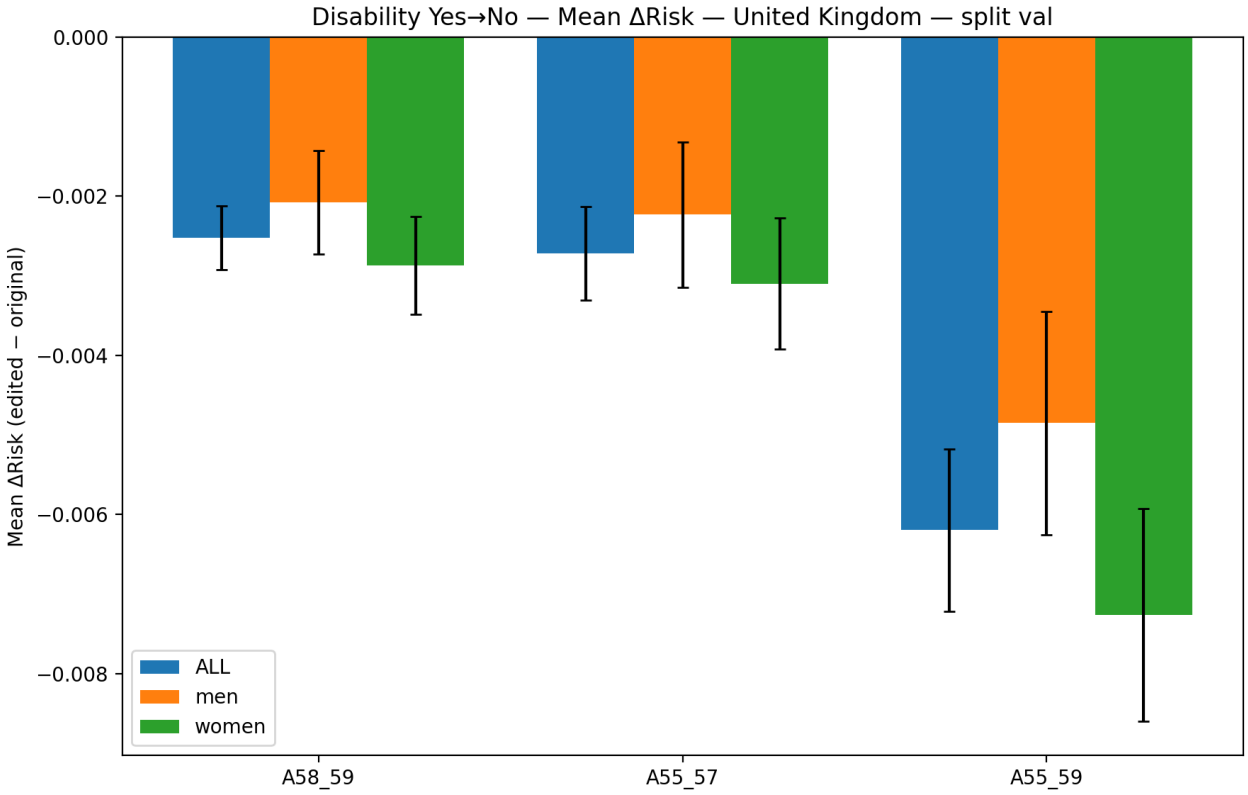


Bars show the mean change in predicted risk (edited − original) when all disability tokens in the indicated age window are flipped from *yes* to *no* in the validation set. Windows: A58–59, A55–57, A55–59. Values are on the probability

scale (0–1). Negative bars mean risk falls after removing disability; positive bars mean risk rises. Error bars depict sampling uncertainty (±1 s.e.). This pooled view shows small average movements overall, with the full-window edit (A55–59) producing the largest absolute shift.

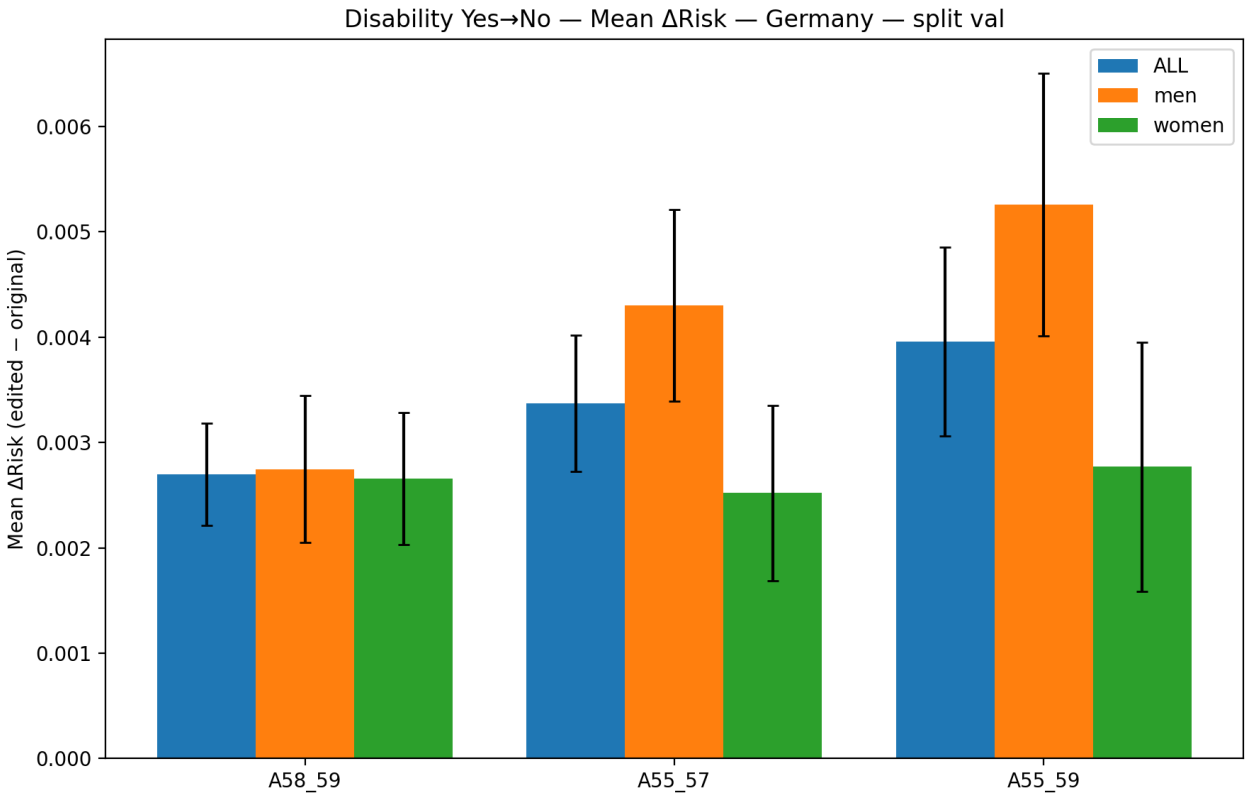
Disaggregating by country clarifies the pattern. In the United Kingdom (Fig. 11), flipping disability to no-disability reduces risk in all windows, with the largest average drop when the edit covers A55–59, and slightly larger absolute reductions for women than for men. In Germany (Fig. 12), by contrast, the same edit raises mean risk slightly in every window; the increase is most pronounced for men and when the edit spans A55–59.

Figure 11. Counterfactual edit (Disability yes→no): mean change in predicted poverty risk - **United Kingdom, by sex**



As in Figure 10, but stratified for the UK and shown separately for ALL, men, and women. All bars are negative, indicating that flipping disability to *no* reduces predicted poverty risk in the UK. The largest average reduction occurs when the edit spans A55–59, with somewhat larger absolute declines for women than for men. Error bars = ±1 s.e.

Figure 12. Counterfactual edit (Disability yes→no): mean change in predicted poverty risk — **Germany, by sex**



As in Figure 10, but for Germany and by sex. Here the bars are positive on average: flipping disability to *no* increases predicted risk slightly, most when the edit covers A55–59, and more for men than for women. This suggests composition/institutional differences in the German data (some late-50s disability records coexist with income-stabilising pathways). Error bars = ±1 s.e.

The UK pattern - risk falling when late-window disability is removed and the largest effects when edits include A55–59 - is consistent with a setting where near-term liquidity and defined-contribution accumulation make late health limitations directly salient for pre-SPA poverty risk. The German pattern - small *increases* in risk after removing disability, strongest among men - suggests selection/composition around who is recorded as disabled late in the 50s: in Germany, disability status near exit can co-occur with continued covered employment (or disability-linked benefits that stabilize income), particularly for men in tenure-intensive occupations. Under that composition, flipping "disabled to no disabled" removes a token that the model has learned to associate with stable pathways for some subgroups, nudging predicted risk up rather than down. This is not a causal claim, but it is a coherent measurement read: the country contrast tracks institutional design (earnings-related accruals and disability pathways in Germany; DC/earnings-liquidity channels in the UK) and gendered job mixes.

These token edits are model-internal counterfactuals: they quantify how the trained classifier's risk assessment moves under stylised changes in late-career histories. They do not identify the causal effect of health improvements or disability programmes.

Together, the edits reinforce the positive-class diagnostics where late-window disability remains a high-leverage signal for screening, but they also surface country- and gender-specific entanglements between disability status, continued covered work, and institutional supports that policymakers should consider when translating risk movements into interventions.

6. Discussion

This study set out to pinpoint when and what in late-career histories carries the most predictive information for near-term poverty. Across diagnostics, a consistent picture emerges. Predictive signal concentrates in the last two pre-outcome years (ages 58–59). For individuals who later become poor, disability is the dominant late-window marker, especially in the United Kingdom, while for those who avoid poverty, working-time stability is the main protective signal, peaking at age 58 in both countries. These patterns are not artifacts of a single metric: they appear in global calibration losses, class-specific ablations, and are directionally corroborated by token-edit experiments.

Differences between countries align with their institutional architectures. In the UK, where living standards before State Pension Age depend heavily on current earnings, liquidity and defined-contribution flows, new or worsening limitations in the late 50s plausibly disrupt work and saving, which is reflected in the strong disability signal among the future poor (Cribb, Hood and Joyce, 2017; Cribb and Emmerson, 2020; Murphy, 2024). In Germany, where entitlements are earnings-related and tied to covered employment, the model places greater weight on continued attachment at the cusp of exit, most clearly for men, while disability is present but less decisive in separating who becomes poor (Börsch-Supan and Ludwig, 2010; OECD, 2019). These country contrasts persist after we hold constant sector and other tokens, suggesting that the mechanisms the model "reads" are consistent with how each system transmits late-career shocks to near-term income risk (OECD, 2010, 2019).

Gender patterns are also coherent with known labour-market profiles. Women show a pronounced response to hours and sector in the negatives-only diagnostics, consistent with higher part-time prevalence and public-service concentration in mid- and late-career (OECD, 2019). Among the positives in the UK, late disability is particularly informative for women, matching their closer proximity to the screening threshold when hours are already reduced (Cribb, Hood and Joyce, 2017; OECD, 2019). Men in Germany exhibit the clearest end-of-window employment signal, consistent with tenure- and strength-intensive jobs where a break in attachment near exit has disproportionate consequences for earnings-related accrual (Börsch-Supan and Ludwig, 2010; OECD, 2019).

The token-edit simulations provide a directional check on these narratives. Flipping disability from "yes" to "no" reduces predicted poverty risk in the UK across windows, with the largest average improvements when the edit spans the full 55–59 horizon and a particularly visible effect for ages 58–59. In Germany, the same edit nudges risk upward on average, small in magnitude but systematic, which is consistent with compositional differences in who is recorded as disabled late in the 50s (including pathways that stabilise income or maintain covered work) and with stronger insurance links to covered employment (OECD, 2010; Börsch-Supan, 2013).

Two policy-relevant messages follow. First, timing: if the objective is to lower near-term poverty risk, the tranche 58–59 is where screening and support are most likely to be informative, with steady monitoring earlier in the window (OECD, 2019). Second, content: in the UK, late-window disability screening and rehabilitation look like promising levers; in Germany, measures that secure attachment and hours, such as short-time work arrangements and retention incentives, target the margin where the model's risk is most sensitive (OECD, 2019; Cribb and Emmerson, 2020; Scarpetta and di Noia, 2023). These are targeting insights, not impact estimates; they are designed to help prioritise where and for whom causal evaluations should be mounted.

Methodologically, treating the Transformer as a diagnostic measurement device added value beyond a black-box predictor. The combination of sliding masks, age×feature ablations, positive-class metrics, and calibrated token edits yielded a policy-readable decomposition by time and mechanism. Rather than a single "importance" score, we recover a compact map: disability at 58–59 is pivotal among the future poor; working-time stability across 55–59 underpins remaining out of poverty; and the relative salience of these margins differs by country and gender in ways that mirror institutional design (Kleinberg *et al.*, 2015; Lundberg and Lee, 2017; Rudin, 2019; Scarpetta and di Noia, 2023). This clarity is precisely what practitioners need to decide whether a one-year push might suffice, whether two–three years are necessary, or whether sustained support throughout the window is warranted.

Finally, the findings set a concrete agenda for evaluation. The heatmaps and edits identify pre-specified ages and levers, late disability mitigation in the UK; end-of-window attachment in Germany; hours stabilisation for women in both countries, where quasi-experimental or experimental designs can most plausibly detect effects (Jacobson, LaLonde and Sullivan, 1993; Chan and Huff Stevens, 2001; OECD, 2019). Read this way, the diagnostic sequence approach is a practical bridge between descriptive prediction and causal testing: it narrows the space, sharpens hypotheses, and makes the timing of late-career interventions empirically tractable.

7. Limitations

Several limitations qualify our findings. First, we do not claim causality. Feature ablations and token-edit experiments are model-internal counterfactuals that probe how the trained classifier reallocates probability mass under stylised edits to late-career histories. Real interventions may trigger selection, behavioural responses, and general-equilibrium effects that our design cannot capture.

Second, measurement differs across ages and countries despite harmonisation. Coverage of some yearly tokens improves with age, most notably disability, so part of the late-window concentration of signal may reflect greater data availability rather than a change in underlying processes. Cross-country asymmetries are substantial: for example, the United Kingdom lacks firm-size information, Germany's ISCO coverage is sparse at earlier ages, and unemployment/activity is fully observed in the UK but not in Germany. These differences can inflate or mute apparent importance and complicate direct comparisons of effect magnitudes.

Third, completeness across the full 55–59 window is imperfect. No individual in the risk set has all yearly features observed at all five ages, and the shortfalls are uneven across variables (e.g.,

disability and working-time are relatively well observed; sector, industry, and ISCO are not). Model-based diagnostics may therefore lean more heavily on features and ages that are better measured, even if the underlying construct is not intrinsically more influential.

Fourth, some predictors are composite "bundles." Working-time status, in particular, packages hours with correlated advantages, health, tenure protection, union coverage, firm size, and job quality. The strong protective signal among negatives should thus be interpreted as the contribution of this bundle rather than of hours alone.

Fifth, threshold-based metrics are sensitive to calibration choices. Our ΔTPR results rely on policy-motivated cut-offs; alternative thresholds would alter magnitudes even if the qualitative age-pattern we document remains.

Finally, results are contingent on the modelling stack (architecture, tokenisation, hyperparameters) and the harmonised data we use. Different encodings or outcome definitions could shift where signal appears. For these reasons, we view the present analysis as a measurement exercise that identifies where and when the model detects policy-relevant information, to be complemented, not replaced, by causal evaluations before drawing conclusions about potential program impact.

Bibliography

Abbott, A. (1995) “Sequence Analysis: New Methods for Old Ideas,” *Annual Review of Sociology*, 21(1), pp. 93–113. Available at: https://doi.org/10.1146/annurev.so.21.080195.000521.

Boeri, T. (2010) “oInstitutional reforms and dualism in European labor markets, pAshenfelter, O. and Card, D.(eds.) Handbook of Labor Economics, vol. 4B, ch. 13.” Amsterdam: Elsevier, p. 1173q236.

Börsch-Supan, A. (2013) “Myths, scientific evidence and economic policy in an aging world,” *The Journal of the Economics of Ageing*, 1, pp. 3–15.

Börsch-Supan, A.H. and Ludwig, A. (2010) *Old Europe ages: reforms and reform backlashes*. National Bureau of Economic Research. Available at: https://www.nber.org/papers/w15744 (Accessed: October 5, 2025).

Breiman, L. (2001) “Random Forests,” *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Chan, S. and Huff Stevens, A. (2001) “Job Loss and Employment Patterns of Older Workers,” *Journal of Labor Economics*, 19(2), pp. 484–521. Available at: https://doi.org/10.1086/319568.

Cribb, J. and Emmerson, C. (2020) “What happens to workplace pension saving when employers are obliged to enrol employees automatically?,” *International Tax and Public Finance*, 27(3), pp. 664–693. Available at: https://doi.org/10.1007/s10797-019-09565-6.

Cribb, J., Hood, A. and Joyce, R. (2017) *Recessions, income inequality and the role of the tax and benefit system*. IFS Report. Available at: https://www.econstor.eu/handle/10419/201778 (Accessed: October 5, 2025).

Dannefer, D. (2003) “Cumulative advantage/disadvantage and the life course: Cross-fertilizing age and social science theory,” *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 58(6), pp. S327–S337.

Davis, J. and Goadrich, M. (2006) “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. *the 23rd international conference*, Pittsburgh, Pennsylvania: ACM Press, pp. 233–240. Available at: https://doi.org/10.1145/1143844.1143874.

Delaney, E., Greene, D. and Keane, M.T. (2021) “Instance-Based Counterfactual Explanations for Time Series Classification,” in A.A. Sánchez-Ruiz and M.W. Floyd (eds.) *Case-Based Reasoning Research and Development*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 32–47. Available at: https://doi.org/10.1007/978-3-030-86957-1_3.

Doshi-Velez, F. and Kim, B. (2017) “Towards A Rigorous Science of Interpretable Machine Learning.” arXiv. Available at: https://doi.org/10.48550/arXiv.1702.08608.

Elder, G.H. (1998) “The Life Course as Developmental Theory,” *Child Development*, 69(1), pp. 1–12. Available at: https://doi.org/10.1111/j.1467-8624.1998.tb06128.x.

Goyal, Y. *et al.* (2019) “Counterfactual visual explanations,” in *International Conference on Machine Learning*. PMLR, pp. 2376–2384. Available at: http://proceedings.mlr.press/v97/goyal19a.html (Accessed: October 5, 2025).

Ismail, A.A., Corrada Bravo, H. and Feizi, S. (2021) “Improving deep learning interpretability by saliency guided training,” *Advances in Neural Information Processing Systems*, 34, pp. 26726–26739.

Jacobson, L.S., LaLonde, R.J. and Sullivan, D.G. (1993) “Earnings losses of displaced workers,” *The American economic review*, pp. 685–709.

Kleinberg, J. *et al.* (2015) “Prediction policy problems,” *American Economic Review*, 105(5), pp. 491–495.

Kuh, D. *et al.* (2003) “Life course epidemiology,” *Journal of epidemiology and community health*, 57(10), p. 778.

Lundberg, S.M. and Lee, S.-I. (2017) “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, 30. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (Accessed: October 5, 2025).

Murphy, E. (2024) “Pensions (Extension of Automatic Enrolment) Bill.” Available at: https://www.niassembly.gov.uk/globalassets/documents/committees/2022-2027/communities/reports/pensions-extension-of-automatic-enrolment-bill/research-papers/20240605-raise---pensions-extension-of-auto-enrolment-bill.pdf (Accessed: October 5, 2025).

OECD (2010) *Sickness, Disability and Work: Breaking the Barriers: A Synthesis of Findings across OECD Countries*. OECD. Available at: https://doi.org/10.1787/9789264088856-en.

OECD (2019) *Working Better With Age*. OECD (Ageing and Employment Policies). Available at: https://doi.org/10.1787/c4d4f66a-en.

Ross, A.S., Hughes, M.C. and Doshi-Velez, F. (2017) “Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations.” arXiv. Available at: https://doi.org/10.48550/arXiv.1703.03717.

Rudin, C. (2019) “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature machine intelligence*, 1(5), pp. 206–215.

Saito, T. and Rehmsmeier, M. (2015) “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PloS one*, 10(3), p. e0118432.

Scarpetta, S. and di Noia, C. (2023) “Pensions at a Glance 2023: OECD AND G20 INDICATORS.,” *Pensions at a Glance* [Preprint]. Available at: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=19954026&asa=N&AN=177979553&h=dWHlZgac6mWk6AMKQ7uf0WGUQbU9w8uaC45C7vdF00kmpd8BZgGmgn9sfYMEO60nzD26flXjtAcz3f40YGv6cw%3D%3D&crl=f (Accessed: October 5, 2025).

Suresh, H. (2017) *Clinical event prediction and understanding with deep neural networks*. PhD Thesis. Massachusetts Institute of Technology. Available at: https://dspace.mit.edu/handle/1721.1/113169 (Accessed: October 5, 2025).

Tonekaboni, S. *et al.* (2019) “What clinicians want: contextualizing explainable machine learning for clinical end use,” in *Machine learning for healthcare conference*. PMLR, pp. 359–380. Available at: https://proceedings.mlr.press/v106/tonekaboni19a.html (Accessed: October 5, 2025).

Turek, K., Voets, I. and Kalmijn, M. (2023) “Comparative Panel File: Manual for CPF v. 1.5.” Available at: https://files.de-1.osf.io/v1/resources/9fhwg_v1/providers/osfstorage/6476f5f3c3a11f04f939ad36?action=download&direct&version=2 (Accessed: October 5, 2025).

Vaswani, A. *et al.* (2017) “Attention is all you need,” *Advances in neural information processing systems*, 30.