

Seeing the Price: Salience, Liquidity Constraints, and Reimbursement-Based Healthcare Design

Evidence and Policy Implications from Japan and Sweden

George Kujo

Independent Researcher and Policy Commentator

ORCID: 0009-0006-6864-0145

Conflict of interest: None declared

Funding: None

Data availability: No original data were generated

Abstract

Healthcare overutilization remains a persistent concern in high-income countries, particularly in systems with low point-of-service cost sharing. Japan exemplifies this issue, recording among the highest outpatient visit rates in the OECD despite achieving health outcomes broadly comparable to Sweden, where utilization is substantially lower. Conventional policy responses have focused on increasing cost-sharing or introducing gatekeeping mechanisms; however, both approaches face substantial political and institutional constraints in the Japanese context.

This paper proposes an alternative framework grounded in the timing and salience of healthcare payments. We argue that healthcare demand is shaped not only by the level of cost-sharing but also by the degree to which costs are visible at the moment of the consumption decision, and by the liquidity constraints patients face when full payment is required upfront. Under conventional in-kind benefit systems, patients observe only a fractional co-payment, potentially inducing moral hazard through reduced price salience.

We introduce a hybrid reimbursement-based model in which patients initially pay the full cost of low-acuity outpatient care—thereby making prices fully salient—while high-cost treatments automatically transition to in-kind coverage to preserve financial protection. This design operationalises the principle of ex-ante discipline combined with ex-post insurance. We further propose integrating digital identity-linked infrastructure to enable real-time reimbursement and individualized cost thresholds, mitigating liquidity constraints and administrative burdens.

Drawing on cross-country comparisons between Japan and Sweden and on causal evidence from studies exploiting cost-sharing discontinuities—most notably Shigeoka (2014)—we argue that substantial reductions in low-value outpatient utilization may be achievable without significant deterioration in population health outcomes. We additionally highlight a novel supply-side implication: improved provider cash flow through reduced reimbursement delays under digitally mediated patient-pay-first arrangements.

The paper contributes to the health economics and health policy literatures by reframing healthcare payment design as a joint problem of price, salience, and financial intermediation. Potential risks—including delayed care-seeking, equity concerns, and implementation challenges—are discussed in detail.

Keywords: healthcare utilization; price salience; reimbursement systems; liquidity constraints; moral hazard; Japan; Sweden; health policy

1. Introduction

High outpatient utilization is one of the defining features of Japan's healthcare system. Japanese residents visit physicians approximately **thirteen times per year** on average—a rate that substantially exceeds the OECD mean of roughly six consultations per capita and surpasses even the highest European comparators.²⁰ Yet despite this intensity of contact, Japan consistently achieves life expectancy among the highest in the world, with age-standardised mortality broadly comparable to Sweden—a country whose outpatient visit rate is substantially lower than Japan's, at fewer than three consultations per capita per year.^{13,20} This apparent paradox raises a fundamental question: is the additional utilization producing health gains, or does it largely represent low-value demand that imposes fiscal cost without commensurate clinical benefit?

The standard economic account attributes high utilization to moral hazard: when patients face low or zero marginal cost at the point of service, they consume more healthcare than they would under full price exposure. The RAND Health Insurance Experiment provided the foundational empirical support for this mechanism, demonstrating that moving from free care to higher cost-sharing substantially reduces utilization.^{1,2} More recent work has confirmed this finding across multiple institutional settings.^{3,4}

Yet the moral hazard framework, as conventionally applied, conflates at least two analytically distinct mechanisms. The first is the standard price effect: patients who pay a lower share of the true cost demand more services. The second is a salience effect: under in-kind benefit systems, patients observe only their co-payment—not the full invoice—rendering the total resource cost invisible at the point of decision. The behavioural economics literature has demonstrated that individuals respond more strongly to salient prices than to equivalent but less visible price signals.^{5,6} If reduced salience independently drives excess utilization, interventions that restore price visibility without raising the long-run patient burden may reduce utilization while preserving access.

A third mechanism—liquidity constraints—complicates the analysis. Even when the long-run insurance contract is generous, some patients cannot produce cash at the moment healthcare is needed. Reimbursement-based systems, under which patients pay upfront and recover costs subsequently, impose a timing mismatch that can discourage care-seeking among cash-constrained individuals.^{7,8} The net effect of a patient-pay-first arrangement therefore depends on the relative magnitudes of the salience effect

(discouraging low-value care) and the liquidity effect (discouraging high-value care among the poor).

This paper develops a framework that engages all three mechanisms jointly. We propose a hybrid reimbursement-based model in which:

- Low-acuity outpatient care is initially paid in full by the patient, making the complete cost salient at the point of service.
- High-cost events—inpatient admissions, specialist procedures, chronic disease management—automatically transfer to in-kind coverage, preserving financial protection where it matters most.
- Digital identity infrastructure enables real-time reimbursement and individualized income-linked thresholds, mitigating liquidity constraints.

We term this design *ex-ante* discipline with *ex-post* insurance, and we argue that it represents a structural solution to the salience–liquidity tension that has limited previous reimbursement-based proposals.

The remainder of the paper is structured as follows. Section 2 reviews the empirical literature on cost-sharing, salience, and liquidity constraints. Section 3 develops the conceptual framework. Section 4 presents the Japan–Sweden comparative case and synthesises the relevant causal evidence. Section 5 specifies the hybrid reimbursement model and its digital infrastructure requirements. Section 6 addresses supply-side implications for provider cash flow. Section 7 discusses equity concerns, potential harms, and implementation risks. Section 8 concludes.

2. Literature Review

2.1 Cost-Sharing, Utilization, and Health Outcomes

The foundational evidence derives from the RAND Health Insurance Experiment (HIE), a large-scale randomised controlled trial conducted in the United States during the 1970s and 1980s.¹ Manning and colleagues demonstrated that participants assigned to free care used substantially more medical services than those facing positive cost-sharing, with no average mortality effect—though low-income and chronically ill participants experienced worse health outcomes under high cost-sharing.¹ Aron-Dine, Einav, and Finkelstein's

re-analysis confirmed the moral hazard finding while emphasising the heterogeneity of health effects and the role of financial frictions in mediating utilization decisions.²

The Oregon Health Insurance Experiment, conducted by Finkelstein and colleagues (2012), provides complementary RCT evidence using a lottery-based expansion of Medicaid eligibility in Oregon.^{17,18} Insurance coverage substantially increased healthcare utilization, yet no statistically significant effect was detected on objective physical health outcomes—including blood glucose, blood pressure, and cholesterol levels—over the study period. Financial protection and mental health improved notably. The convergence of the RAND and Oregon findings across two independent RCTs in different eras and populations reinforces the interpretation that marginal increases in utilization induced by lower cost exposure do not reliably translate into gains in physical health outcomes, lending further empirical support to the premise that a non-trivial share of utilization at the margin is of low clinical value.

Einav and Finkelstein's synthesis concluded that more generous insurance consistently increases spending, with the welfare implications of cost-sharing reforms depending heavily on whether the suppressed utilization is high-value or low-value.³ Blunt increases in patient cost exposure risk disproportionately deterring necessary care among vulnerable populations.

In the Japanese context, Shigeoka's (2014) regression discontinuity analysis is the most directly relevant causal evidence.⁴ Exploiting the sharp reduction in cost-sharing at age 70 under Japan's elderly healthcare law, Shigeoka documented substantial increases in outpatient and inpatient utilization at the threshold, with little detectable effect on mortality over a five-year window. This supports two conclusions: Japanese patients are price-sensitive, and the additional utilization induced by lower cost-sharing did not, on average, produce health gains—consistent with a substantial proportion of low-value demand at the margin.

2.2 Liquidity Constraints and Payment Timing

Zhong (2011) analysed variation in reimbursement method across Chinese healthcare settings and found that patient-pay-first arrangements reduced outpatient utilization relative to direct third-party payment, identifying liquidity constraints and transaction costs as plausible mechanisms.⁷

Belchior and Gomes (2022) provided direct evidence that liquidity suppresses healthcare demand, exploiting Brazil's COVID-19 Emergency Aid (Auxílio Emergencial) cash transfer program as a natural experiment. Households receiving the transfer increased healthcare use, particularly among low-income groups, implying that prior unmet demand partly reflected inability to pay at the moment of service.⁸

Gross, Layton, and Prinz (2022) documented pronounced liquidity sensitivity in prescription drug filling and other health spending among older American adults, with spending spiking immediately after Social Security payments.⁹ Lyngse (2020) found similar payment-date effects in Scandinavian administrative data.¹⁰ These studies establish that any reimbursement-based system requiring upfront payment must account for the distributional incidence of liquidity constraints, which fall disproportionately on lower-income, older, and chronically ill populations.

2.3 Salience and Behavioural Mechanisms

Baicker, Mullainathan, and Schwartzstein (2015) developed a formal model of *behavioural hazard* in health insurance, distinguishing between moral hazard driven by reduced price and excess utilization driven by present bias and inattention.⁵ Their framework implies that making costs more visible—raising salience—may achieve utilization reductions without the access harms associated with blunt price increases.

Handel and Kolstad (2015) demonstrated the substantive role of information frictions and hassle costs in health insurance plan choice, showing that these frictions have meaningful implications for consumer welfare. Building on this, Domurat, Menashe, and Yin (2021) provided direct field-experimental evidence: a randomized intervention mailed to approximately 87,000 households in California's health insurance marketplace showed that personalized, salience-enhancing information materially affected enrollment behavior, substantiating the hypothesis that salience is a quantitatively important and operationally manipulable design variable.^{6,19} Abaluck and Gruber (2011) documented systematic decision-making errors among Medicare Part D enrollees, further motivating the case for choice architecture and automated decision support in digital reimbursement systems.¹¹

2.4 Cross-Country Comparisons: Japan and Sweden

The OECD's assessment of the Japanese healthcare system identified fee-for-service reimbursement, low patient cost exposure, and unrestricted direct access to specialists as structural contributors to high outpatient utilization.¹²

Sato and colleagues (2024) provided a Japan–Sweden comparison of healthcare systems, documenting that Japan's physician consultation rate far exceeds Sweden's despite broadly comparable life expectancy.¹³ The OECD primary care report (2020) associated strong primary care gatekeeping with fewer avoidable hospitalizations and better chronic disease management.¹⁴ The Japan–Sweden contrast illustrates a case where substantially lower utilization coexists with equivalent aggregate health outcomes, suggesting significant scope for utilization reduction in Japan without population health harm, subject to important distributional caveats.

2.5 Digital Infrastructure and Reimbursement Automation

The WHO Regional Office for Europe (2023) identified timely and predictable reimbursement as a key determinant of digital health tool adoption, arguing that structured digital reimbursement pathways reduce uncertainty for both providers and patients.¹⁵ Japan's existing MyNumber digital identity infrastructure and electronic medical record systems provide a plausible technological foundation for the model proposed in this paper, a point we develop further in Section 5.

3. Conceptual Framework

3.1 Decomposing the Demand for Healthcare

Standard models of healthcare demand under insurance treat the relevant price as the co-insurance rate multiplied by the full cost of care: reducing the co-insurance rate lowers the effective price and raises quantity demanded. This is the conventional moral hazard mechanism.

We propose a more granular decomposition. Let the quantity of healthcare consumed by individual i for encounter type j be:

$$Q_{ij} = f(P_{ij}, S_{ij}, L_i, X_i)$$

where P is the effective price (co-payment), S is the salience of the full cost at the point of service, L is the degree of liquidity constraint, and X is a vector of health status and preference covariates. The key addition is S and L as independent arguments.

Under a pure in-kind system, S is low because the patient observes only the co-payment, not the full invoice. Under a reimbursement-based system in which the patient pays the full fee upfront and receives a detailed receipt before claim submission, S is high by design. The salience effect operates through the difference in S holding P constant.

The liquidity effect operates through L : patients with low liquidity are more sensitive to the upfront payment requirement and may defer or avoid care irrespective of their long-run insurance generosity. This effect is orthogonal to price and salience in that it acts through the timing of cash flows rather than cost perception.

3.2 In-Kind Benefits and the Perception of Price

Under Japan's current system, a patient visiting an outpatient clinic pays a co-payment of 10–30% of the fee-schedule price, with the remainder settled directly between insurer and provider. From the patient's perspective, the relevant price signal at the moment of decision is the co-payment, not the full cost. In the case of a consultation billed at ¥3,000, a patient with 10% co-insurance pays ¥300 and may be entirely unaware that the insurer paid the remaining ¥2,700.

This structure systematically understates the resource cost of care in the patient's decision calculus. The behavioural economics literature on salience suggests that such understatement is not merely a notational artefact: individuals genuinely respond more to visible prices than to equivalent but less visible signals.^{5,6} A patient who observes the full ¥3,000 charge before submitting a claim may apply a stricter implicit threshold for judging whether a visit is worthwhile—even if they anticipate recovering most of the fee subsequently. We term this difference the salience-induced component of excess utilization.

3.3 Reimbursement Systems as a Self-Selection Mechanism

An important implication of the salience framework is that a patient-pay-first system functions, in part, as a behavioural self-selection mechanism. Patients who assess a visit as genuinely necessary will proceed despite the upfront payment and the administrative friction of claiming reimbursement. Patients uncertain about the value of a visit may face a higher implicit barrier—not because of the long-run price but because of the salience of the full cost and the perceived effort of the reimbursement process.

This self-selection channel is normatively ambiguous. If the deterred visits are predominantly low-value, the mechanism achieves efficiency gains without health harm—consistent with Shigeoka's (2014) finding that marginal utilization at Japan's age-70 threshold had little detectable health effect.⁴ If, however, deterred visits include necessary care that patients cannot evaluate accurately, the mechanism risks suppressing high-value utilization, particularly among patients with limited health literacy or high liquidity constraints. The hybrid model in Section 5 exploits the self-selection channel for low-acuity care while insulating high-value care through automatic transition to in-kind coverage.

3.4 Ex-Ante Discipline and Ex-Post Insurance

We formalise the design objective as follows. An efficient reimbursement system should impose ex-ante discipline—through salience, transaction costs, and explicit price observation—on low-value utilization decisions, while providing ex-post insurance—through full or near-full reimbursement—against the financial consequences of high-cost events. This is structurally analogous to deductible-based insurance in other sectors but differs in its emphasis on salience as the primary mechanism rather than long-run cost exposure.

The ex-ante/ex-post distinction has an important temporal dimension. Under conventional insurance, the patient makes utilization decisions under conditions of reduced price salience, and financial protection is provided simultaneously at the point of service. Under our proposed model, the patient decides under conditions of high salience, and insurance protection is delivered ex-post through rapid reimbursement. The net financial outcome may be similar across systems, but the timing and visibility of the cost signal differ substantially.

4. Empirical Grounding: Japan and Sweden

4.1 Japan's Utilization Profile

Japan records among the highest outpatient visit rates in the OECD, exceeding twelve consultations per capita per year.^{12,20} This utilization intensity has persisted across multiple decades and financing reforms, suggesting structural rather than transient origins. Japan's system combines universal public insurance with relatively low

co-payment rates and unrestricted direct access to both general practitioners and hospital-based specialists.

The fee-for-service structure creates provider incentives that complement patient demand for frequent contact. Japanese hospitals, which operate both inpatient and extensive outpatient departments, have historically competed for outpatient volume. Per-visit fees have been adjusted through the national fee schedule to partially counteract this tendency without eliminating the underlying incentive structure.

Despite this utilization intensity, Japan achieves life expectancy substantially exceeding the OECD average with per-capita health expenditure that remains moderate relative to GDP and substantially below that of the United States. The combination of high utilization and moderate expenditure is partly attributable to Japan's low fee-schedule prices, but the aggregate resource volume nonetheless raises questions of allocative efficiency.¹²

4.2 Sweden as a Comparator

Sweden's healthcare system differs architecturally in several important respects. Primary care physicians serve as gatekeepers to specialist services, and direct access to hospital outpatient departments is restricted except in emergency settings.^{13,14} Patient fees are regulated and structured to encourage primary care contact rather than specialist self-referral.

Sweden's outpatient consultation rate is approximately 2.9 physician visits per capita per year—less than a quarter of Japan's figure—yet life expectancy and most health outcome indicators are broadly comparable.¹³ This comparison does not establish causality; confounding by demographics, disease prevalence, and primary care quality is substantial. It does, however, constitute a proof of concept: population-level health outcomes need not deteriorate under a system architecture generating substantially fewer outpatient contacts (see Figure 1).

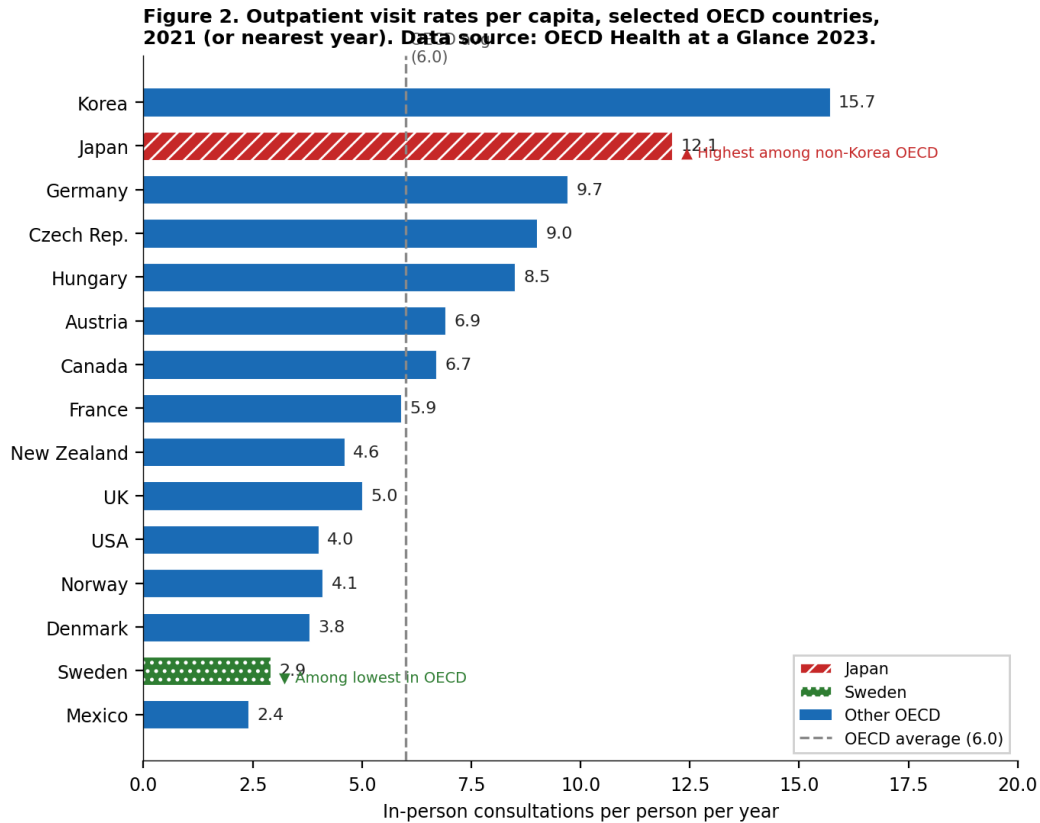


Figure 1. Outpatient visit rates per capita, selected OECD countries, 2021 (or nearest year). Japan and Sweden annotated. Data source: OECD Health at a Glance 2023 (doi:10.1787/7a7afb35-en).

4.3 Causal Evidence: Shigeoka (2014)

Shigeoka (2014) provides the most credible causal evidence on price sensitivity in the Japanese outpatient context.⁴ The paper exploits a sharp discontinuity in cost-sharing at age 70 under Japan's elderly healthcare law: patients aged 70 and above faced a 10% co-insurance rate, compared with 20–30% for younger age groups. At the threshold, Shigeoka documented a significant and immediate increase in both outpatient and inpatient utilization, consistent with strong price sensitivity.

Two findings are particularly relevant. First, the estimated price sensitivity confirms that Japanese patients respond to changes in observed out-of-pocket cost, establishing a plausible mechanism for salience-based interventions. Second, Shigeoka found no statistically significant effect on mortality over a five-year window,⁴ suggesting that additional utilization induced by lower cost-sharing did not produce commensurate health gains at the margin—consistent with a non-trivial share of Japanese outpatient visits being of low marginal clinical value.

This finding does not imply that utilization reduction is harmless for all patients or conditions. The mortality outcome is a blunt measure that may miss morbidity effects, delayed diagnosis, and quality-of-life impacts. The null result applies to the population near the age-70 threshold, which may differ from younger or poorer populations in clinically important ways. We return to these heterogeneity concerns in Section 7.

4.4 OECD-Level Patterns

At the cross-national level, OECD data reveal substantial variation in outpatient utilization rates that is not explained by health outcome differences.^{12,14,20} Countries with stronger primary care orientation—including the Netherlands, Sweden, Denmark, and the United Kingdom—consistently report lower specialist consultation rates and fewer avoidable hospitalizations, while maintaining competitive life expectancy statistics.¹⁴ This pattern supports the hypothesis that a portion of the utilization gap between Japan and these comparators represents low-value demand, though the mechanisms remain contested.

5. The Hybrid Reimbursement Model

5.1 Model Architecture

The hybrid reimbursement model is structured around three interacting components: a salience-maximising payment layer for low-acuity outpatient care; an automatic transition rule to in-kind coverage for high-cost events; and a digital identity-linked infrastructure enabling real-time reimbursement and individualized thresholds.

Under the base design, all outpatient consultations and low-acuity services below a specified cost threshold are subject to full upfront patient payment. The patient receives a detailed invoice displaying both the total fee-schedule cost and the expected reimbursement amount, ensuring that the full resource cost is visible before the encounter is completed. Reimbursement is processed digitally within a specified window to the patient's designated account.

For services above the cost threshold—inpatient admissions, high-cost diagnostics, surgical procedures, and long-term condition management—the insurer assumes direct liability, reverting to an in-kind benefit structure. Patients with confirmed chronic diagnoses or predicted annual expenditures exceeding a specified income-percentile

threshold may receive blanket in-kind coverage for condition-related services regardless of individual encounter costs (see Figure 2).

Figure 1. Conceptual diagram of the hybrid reimbursement model



Figure 2. Conceptual diagram of the hybrid reimbursement model. The figure illustrates the decision pathway for a patient seeking outpatient care: (a) patient pays full fee upfront; (b) full cost made salient via itemized receipt; (c) digital claim submitted in real time by provider; (d) insurer processes claim; reimbursement credited to patient within three to seven business days (deliberate delay to preserve salience); (e) net out-of-pocket cost equals pre-specified co-insurance rate. For encounters above the cost threshold, the pathway transitions to direct in-kind coverage (dashed line).

5.2 Threshold Design and Individualization

A critical design parameter is the cost threshold at which the system transitions from patient-pay-first to in-kind coverage. Setting the threshold too low limits the salience effect to a narrow band of services; setting it too high exposes patients to financially significant upfront payments that may deter necessary care among liquidity-constrained individuals.

We propose that the threshold be individualized using income data held by the tax authority and health expenditure data held by the insurer, both accessible via the digital identity system. A low-income patient might face a threshold of ¥3,000 per encounter (approximately the cost of a standard outpatient consultation), above which in-kind coverage activates. A higher-income patient might face a threshold of ¥10,000–15,000. This structure ensures that the salience effect is proportionate to income, minimizing access barriers for lower-income groups.

Annual expenditure caps, analogous to out-of-pocket maxima in other systems, provide a backstop against catastrophic cumulative liability. Patients approaching their annual cap automatically transition to full in-kind coverage for the remainder of the calendar year (see Figure 3).

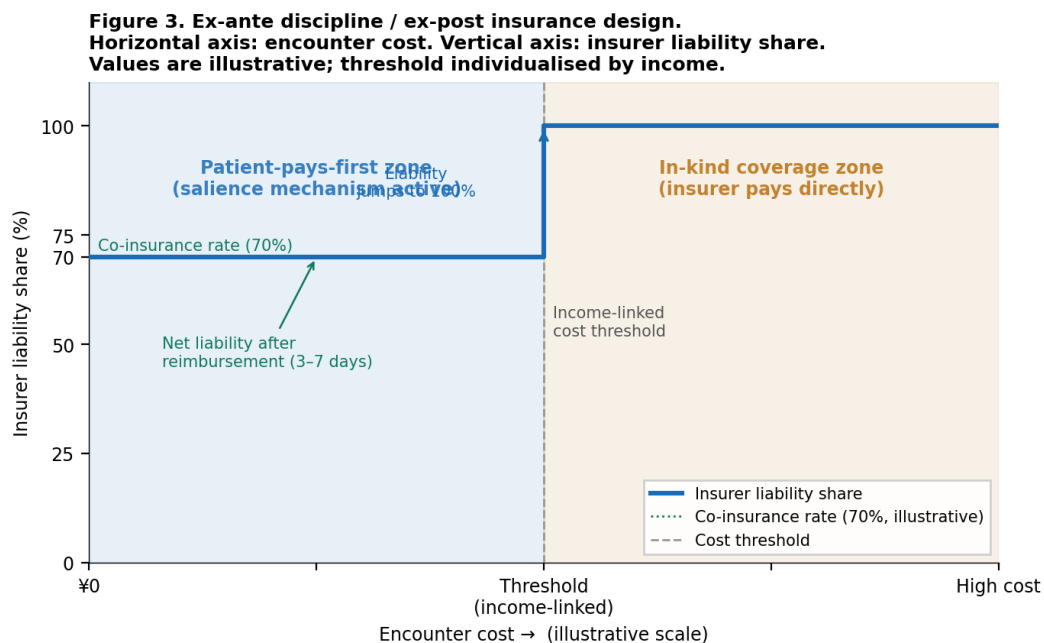


Figure 3. Ex-ante discipline / ex-post insurance design. Horizontal axis: encounter cost (illustrative scale). Vertical axis: insurer liability share (%). Below the income-linked cost threshold, patient liability is 100% upfront; rapid reimbursement reduces net liability to the co-insurance rate. Above the threshold, insurer liability reverts to 100% (in-kind) immediately. Values are illustrative; threshold is individualised by income.

5.3 Digital Identity Infrastructure and Real-Time Reimbursement

The operational feasibility of the hybrid model depends critically on the speed and reliability of the reimbursement process. If patients must wait weeks for reimbursement, the liquidity constraint effect may dominate the salience effect, deterring care-seeking regardless of the long-run financial outcome. Digital identity infrastructure resolves this tension by enabling near-instantaneous claim processing.

Japan's MyNumber system, introduced in 2016 and progressively integrated with health insurance records and medical information systems, provides a foundational infrastructure for this purpose. Upon completing a consultation, the provider would submit a digital claim in real time to the insurer, which would process and approve reimbursement for administrative record-keeping purposes. A critical design question concerns the optimal reimbursement delay. If reimbursement reaches the patient on the same day or the following business day, the cash-flow gap is negligible—which maximises financial protection and equity, but may attenuate the salience effect: a patient who knows they will be reimbursed almost immediately may behave similarly to one under a conventional in-kind system, undermining the behavioural discipline that motivates the design. Conversely, an excessively long delay restores the full liquidity burden and risks deterring necessary care among constrained patients, reproducing the access harms documented in the cost-sharing literature. We therefore propose a deliberately calibrated intermediate delay of three to seven business days as the baseline reimbursement window. This interval is long enough to preserve the salience effect—the patient must consciously observe, record, and await recovery of the full upfront cost—yet short enough to avoid imposing genuine financial hardship on most households. The optimal window should be treated as an empirical parameter, subject to calibration through pilot programmes monitoring both utilization rates and care-avoidance indicators across income groups. For patients with documented low liquidity or chronic conditions qualifying for priority status, an accelerated one-business-day window or provisional in-kind coverage pending reimbursement could be offered, ensuring that the deliberate delay does not fall inequitably on the most financially vulnerable.

This rapid reimbursement architecture transforms the patient-pay-first arrangement from a liquidity barrier into a pure salience mechanism. For patients with very low liquidity, an option for provisional in-kind coverage pending reimbursement—analogueous to a

micro-credit facility—could be incorporated to ensure that the system does not systematically exclude the most financially vulnerable.

5.4 Scope of Application and Service Classification

The model is designed for application to ambulatory outpatient care—consultations, routine diagnostics, prescription renewal visits, and similar low-acuity encounters—rather than to emergency, inpatient, or high-complexity services. This scope restriction reflects the concentration of low-value utilization in the outpatient sector.

Service classification would use Japan's existing fee schedule taxonomy. A schedule of encounter types eligible for patient-pay-first treatment would be established by the relevant authority, updated periodically, and communicated through the digital infrastructure. Services for which clinical guidelines indicate high-value routine use—vaccination, cancer screening, maternal care—would be explicitly exempted regardless of cost.

6. Supply-Side Implications: Provider Cash Flow

The literature on healthcare payment design has focused predominantly on demand-side effects. A less-examined implication concerns the supply side: how providers are affected by changes in the timing and source of payment flows.

Under Japan's current insurance system, providers submit claims and typically receive payment within one to two months of service delivery. This reimbursement lag creates a working capital requirement, particularly for smaller clinics that lack the liquidity reserves of hospital systems. Administrative rejection rates introduce further uncertainty into provider cash flows.

Under a patient-pay-first system, providers receive immediate full payment at the time of the encounter, eliminating the reimbursement lag and the associated working capital cost. The insurer's obligation shifts from a direct payment to the provider to a reimbursement to the patient, simplifying the insurer-provider billing relationship.

This supply-side benefit is particularly significant for small independent clinics, which constitute a large share of Japan's outpatient sector. Improved provider cash flow may increase financial stability and reduce administrative burden, with potential secondary

effects on practice viability in rural and underserved areas. It may also reduce incentives for claim inflation, since the direct billing relationship with insurers is replaced by transparent fee collection at the point of service.

A potential countervailing concern is that providers might set fees above the regulated fee-schedule price if insurer monitoring is diminished. The model should therefore retain regulatory price controls and audit mechanisms. The digital infrastructure required for real-time reimbursement also supports real-time fee transparency: patients would be able to verify the declared fee against the fee schedule before completing payment, creating a decentralized price-checking mechanism.

7. Equity, Potential Harms, and Implementation Challenges

7.1 Equity Concerns

The most serious potential harm from a patient-pay-first system is the risk of delayed or forgone care among populations with low liquidity, poor health literacy, or high clinical need. Existing evidence documents that harm from cost-sharing is not distributed equally: lower-income, older, and sicker patients are disproportionately likely to reduce clinically necessary care in response to upfront payment requirements.¹⁶

Our model addresses this risk through three mechanisms. First, income-linked individualization of the cost threshold limits financial exposure of lower-income patients. Second, real-time digital reimbursement minimizes the liquidity gap between payment and recovery. Third, automatic in-kind coverage for chronic conditions and high-expenditure patients insulates the most clinically vulnerable from the patient-pay-first layer.

Residual equity risks remain. Patients without reliable bank account access or digital connectivity may face barriers to real-time reimbursement. Health literacy constraints may impair accurate self-assessment of clinical need, leading low-health-literacy patients to under-utilise care even when their conditions are serious. Implementation should be accompanied by sustained patient education, accessible claim submission channels, and monitoring of utilization patterns stratified by income, age, and chronic condition status.

7.2 Risk of Delayed Diagnosis and Care Escalation

A concern specific to salience-based utilization reduction is that deterred visits may include early-stage consultations that, if avoided, allow conditions to progress to more severe and costly stages. If patients with early cancer symptoms, uncontrolled hypertension, or evolving infectious illness are dissuaded by upfront payment requirements, the system may reduce low-acuity utilization at the cost of increased downstream inpatient costs and health deterioration.

The evidence from Shigeoka (2014) does not support large average mortality effects from reduced utilization at the margin of Japan's age-70 threshold.⁴ However, this result applies to a specific population and cannot be straightforwardly extrapolated to a system-wide reimbursement reform. Ongoing outcome monitoring—including tracking of hospitalization rates for ambulatory-care-sensitive conditions—would be an essential component of any implementation programme.

7.3 Implementation Challenges

Several implementation challenges deserve explicit recognition. First, public acceptance of a shift from in-kind to reimbursement-based payment may be limited, particularly among older populations accustomed to minimal upfront cost. A phased introduction—beginning with voluntary enrolment among higher-income groups—may be politically more feasible than system-wide transition.

Second, the digital infrastructure required for real-time reimbursement, while technically feasible given Japan's MyNumber system, requires substantial investment in interoperability between health insurance records, medical provider systems, and financial payment infrastructure. Data privacy and security concerns around the consolidation of health and financial data must be addressed through robust regulatory frameworks and patient consent mechanisms.

Third, the service classification process requires clinical input and regular updating as the evidence base evolves. The risk of classification error—in which high-value services are inadvertently subjected to upfront payment requirements—must be managed through conservative initial scope and active post-implementation review.

8. Conclusion

This paper has argued that healthcare utilization in high-income systems is shaped not only by the level of patient cost-sharing but also by the salience of costs at the point of decision and by the liquidity constraints affecting patients' ability to produce cash at the moment of service. Conventional moral hazard analyses have conflated these mechanisms, limiting the range of policy tools considered.

We have proposed a hybrid reimbursement model—combining ex-ante discipline through full-cost visibility with ex-post insurance through rapid digital reimbursement—as a theoretically coherent and operationally feasible alternative to blunt cost-sharing increases or gatekeeping reforms. The Japan–Sweden comparative case, combined with causal evidence from Shigeoka (2014) and the broader literature on liquidity constraints and behavioural salience, provides empirical grounding for the argument that substantial reductions in low-value outpatient utilization may be achievable without commensurate deterioration in population health outcomes.

The model is not without risks. Equity concerns, the possibility of delayed care-seeking, and the substantial implementation requirements of real-time digital reimbursement infrastructure all deserve serious attention. We have proposed design features to mitigate each risk—income-linked thresholds, automatic in-kind transitions, and digital provisional credit—and acknowledge that their effectiveness depends on implementation quality and sustained monitoring.

Three novel contributions emerge from this analysis. First, we offer a salience-based reinterpretation of moral hazard in Japanese healthcare that identifies a utilization-reduction lever distinct from price. Second, we introduce the hybrid ex-ante/ex-post design as a means of reconciling salience-based discipline with financial protection. Third, we highlight the supply-side cash-flow implications of patient-pay-first systems as an underexplored benefit for provider financial stability.

Future research should focus on empirical estimation of the salience effect separate from the price effect in quasi-experimental designs, on pilot evaluation of rapid digital reimbursement systems in controlled settings, and on the development of validated instruments for classifying outpatient services by marginal clinical value.

Figure Captions

[Figures are embedded in the text above.]

References

1. Manning WG, Newhouse JP, Duan N, Keeler EB, Leibowitz A, Marquis MS. Health insurance and the demand for medical care: evidence from a randomized experiment. *Am Econ Rev.* 1987;77(3):251–277.
2. Aron-Dine A, Einav L, Finkelstein A. The RAND Health Insurance Experiment, three decades later. *J Econ Perspect.* 2013;27(1):197–222. doi:10.1257/jep.27.1.197
3. Einav L, Finkelstein A. Moral hazard in health insurance: what we know and how we know it. *J Eur Econ Assoc.* 2018;16(4):957–982. doi:10.1093/jeea/jvy017
4. Shigeoka H. The effect of patient cost sharing on utilization, health, and risk protection. *Am Econ Rev.* 2014;104(7):2152–2184. doi:10.1257/aer.104.7.2152
5. Baicker K, Mullainathan S, Schwartzstein J. Behavioral hazard in health insurance. *Q J Econ.* 2015;130(4):1623–1667. doi:10.1093/qje/qjv029
6. Handel BR, Kolstad JT. Health insurance for "Humans": information frictions, plan choice, and consumer welfare. *Am Econ Rev.* 2015;105(8):2449–2500. doi:10.1257/aer.20131126
7. Zhong H. Effect of patient reimbursement method on health-care utilization: evidence from China. *Health Econ.* 2011;20(11):1312–1329. doi:10.1002/hecl.1670
8. Belchior CA, Gomes Y. Liquidity constraints, cash transfers and the demand for health care in the Covid-19 pandemic. *Health Econ.* 2022;31(11):2369–2380. doi:10.1002/hecl.4585
9. Gross T, Layton TJ, Prinz D. The liquidity sensitivity of healthcare consumption: evidence from Social Security payments. *Am Econ Rev Insights.* 2022;4(2):175–190. doi:10.1257/aeri.20200830
10. Lyngse FP. Liquidity Constraints and Demand for Healthcare: Evidence from Danish Welfare Recipients. arXiv preprint arXiv:2010.14651. 2020. doi:10.48550/arXiv.2010.14651
11. Abaluck J, Gruber J. Choice inconsistencies among the elderly: evidence from plan choice in the Medicare Part D program. *Am Econ Rev.* 2011;101(4):1180–1210. doi:10.1257/aer.101.4.1180. PMID:21857716
12. Jeong HS, Hurst J. An assessment of the performance of the Japanese health care system. OECD Labour Market and Social Policy Occasional Papers, No. 56. Paris: OECD Publishing; 2001. doi:10.1787/445332508603
13. Sato R, Jakobsson U, Midlöv P. A proposed medical system change in Japan inspired by Swedish primary health care: important role of general practitioners and specialist nurses at primary health care centers. *J Gen Fam Med.* 2024;25(6):295–304. doi:10.1002/jgf2.726

14. OECD. Realising the potential of primary health care. Paris: OECD Publishing; 2020. doi:10.1787/a92adee4-en
15. World Health Organization Regional Office for Europe. Digital health in the European Region: the ongoing journey to commitment and transformation. Copenhagen: WHO Regional Office for Europe; 2023.
16. Rice T, Matsuoka KY. The impact of cost-sharing on appropriate utilization and health status: a review of the literature on seniors. *Med Care Res Rev.* 2004;61(4):415–452. doi:10.1177/1077558704269498
17. Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, et al. The Oregon Health Insurance Experiment: evidence from the first year. *Q J Econ.* 2012;127(3):1057–1106. doi:10.1093/qje/qjs020. PMID:23293397
18. Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, et al. The Oregon experiment — effects of Medicaid on clinical outcomes. *N Engl J Med.* 2013;368:1713–1722. doi:10.1056/NEJMs1212321
19. Domurat R, Menashe I, Yin W. The role of behavioral frictions in health insurance marketplace enrollment and risk: evidence from a field experiment. *Am Econ Rev.* 2021;111(5):1549–1574. doi:10.1257/aer.20190823
20. OECD. Health at a Glance 2023: OECD Indicators. Paris: OECD Publishing; 2023. doi:10.1787/7a7afb35-en