



UNIVERSITY
OF WARSAW



FACULTY OF
ECONOMIC SCIENCES

WORKING PAPERS

No. 6/2026 (500)

COST-EFFECTIVENESS, INSURANCE SOLVENCY, AND HEALTH CAPITAL: A JOINT INCOMPATIBILITY THEOREM AND DYNAMIC RESEARCH PROGRAMME

JAKUB RYŁOW

WARSAW 2026

ISSN 2957-0506



Cost-Effectiveness, Insurance Solvency, and Health Capital: A Joint Incompatibility Theorem and Dynamic Research Programme

*Jakub Rylow¹**

¹ *Faculty of Economic Sciences, University of Warsaw*

* *Corresponding author: jakub.jan.rylow@gmail.com*

Abstract: This paper examines the mathematical architecture of health economics through four analytical traditions—Arrow’s welfare economics of medical care, Grossman’s health capital model, Garber’s cost-effectiveness analysis, and Zweifel’s actuarial insurance theory—and derives three original results. C1 (Joint Incompatibility Theorem): under empirically standard regularity conditions, no health system simultaneously satisfies the Garber–Phelps WTP-based cost-effectiveness criterion, Zweifel’s actuarial solvency condition under community rating, and the Pauly–Zweifel optimal coinsurance rule; the proof proceeds via three lemmas, each isolating a separate incompatibility. C2 (Dynamic Cost-Effectiveness Threshold): the cost-effectiveness threshold $\lambda^*(t)$ is derived as an endogenous co-state variable of a unified dynamic optimisation, and shown to equal the present value of future marginal health benefits discounted at the user cost of health capital $\rho + \delta(t)$. C3 (Demographic Suboptimality of Uniform Thresholds): as a corollary, uniform thresholds are dynamically inconsistent with any population whose age-specific depreciation rate is increasing, with the direction of the depreciation channel being theoretically determinate. Together, C1 converts a collection of separately noted tensions into a single structural result, while C2 and C3 convert that result into a constructive research programme for dynamic health economic evaluation.

Keywords: health economics, cost-effectiveness threshold, health capital, actuarial solvency, moral hazard, dynamic cost-effectiveness, QALY, Grossman model, incompatibility theorem, endogenous threshold

JEL codes: D81, I11, I13, C61, D82, G22

1 Introduction

The economics of medicine occupies an unusual position within the broader discipline. It is simultaneously a field of applied microeconomics, a branch of welfare theory, and a domain in which the mathematical formalism of decision theory, actuarial science, and statistical inference intersect with questions of fundamental social importance. Unlike commodity markets, health care markets are constitutively structured by uncertainty—about the onset of illness, the efficacy of treatment, the quality of information available to patients and providers, and the long-run fiscal sustainability of insurance arrangements. Understanding these features requires not only empirical investigation but rigorous theoretical foundations.

This paper examines three bodies of work—those of Kenneth Arrow, Alan Garber, and Peter Zweifel—that have shaped those foundations. This selection requires justification. Health economics has other canonical contributors: Michael Grossman’s health capital model (Grossman, 1972) formalised the *demand* for health as distinct from the demand for medical care; Alain Enthoven’s theory of managed competition shaped actual health system design in the United States and beyond; Anthony Culyer’s need-based allocation framework provides the primary normative competitor to QALY-based utilitarianism. The choice of Arrow, Garber, and Zweifel is not a claim that they exhaust the field. It is a claim that they represent three analytically distinct mathematical traditions—welfare economics under uncertainty, decision-theoretic resource allocation, and actuarial insurance theory—whose relationship to one another has not been examined with sufficient precision. Grossman is integrated into the analysis below precisely because his framework is indispensable for understanding the demand-side complement to Arrow’s supply-side diagnosis.

The central argument this paper makes is *not* that these three traditions converge harmoniously. They do not. They produce mutually incompatible policy prescriptions when applied simultaneously to the same health system: the welfare-maximising cost-effectiveness threshold implied by Garber and Phelps, the actuarially solvent premium structure required by Zweifel, and the optimal coinsurance rate derived from the moral hazard literature cannot all be implemented without violating at least one framework’s foundational conditions (Theorem 1). But identifying incompatibility precisely is not the end of the analysis — it is its starting point. The incompatibility result reveals the structure that a genuinely unified health economics would require: a framework in which the cost-effectiveness threshold is not an exogenous policy parameter but an endogenous quantity determined jointly by health capital dynamics, actuarial solvency conditions, and consumer incentive constraints. Section 8 sketches this framework and derives its central object — the dynamic cost-effectiveness threshold $\lambda^*(t)$ — as a forward-looking integral over demographic and fiscal dynamics. The paper thereby moves from diagnosis to programme.

Original contributions. Beyond its synthetic function, the paper makes three analytically original contributions.

C1 (Joint Incompatibility Theorem, Section 7). We formally prove that the three policy frameworks are mutually incompatible under empirically plausible conditions. The result is not a qualitative observation that tensions exist — those have been noted before — but a theorem with an explicit proof establishing that the set of parameter configurations satisfying all three frameworks simultaneously is empirically empty (Theorem 1, Corollary 1). The proof proceeds via three lemmas, each deriving a contradiction from one of three regularity assumptions: diminishing marginal utility of income, strictly increasing age-specific health capital depreciation, and elastic demand for medical care.

C2 (Dynamic Cost-Effectiveness Threshold, Section 8). We derive the cost-effectiveness threshold $\lambda^*(t)$ as an endogenous co-state variable of a unified dynamic optimisation that jointly embeds the Garber–Phelps allocation rule, Grossman’s health capital dynamics, and Zweifel’s actuarial solvency constraint (Proposition 1). The result— $\lambda^*(t)$ equals the present value of future marginal health benefits discounted at the user cost of health capital $\rho + \delta(t)$ —provides the first formal characterisation of the threshold as a function of demographic dynamics rather than a policy constant. This constitutes the theoretical foundation for a dynamic theory of cost-effectiveness that the literature has identified as a goal but has not constructed.

C3 (Demographic Suboptimality of Uniform Thresholds, Section 8). We show as a corollary of C2 that uniform cost-effectiveness thresholds are dynamically inconsistent with any population whose age composition is changing (Corollary 2). The direction of the bias is theoretically determinate: in ageing populations, the optimal threshold $\lambda^*(t)$ falls as $\delta(t)$ rises, so static thresholds systematically overstate the cost-effectiveness of health investments relative to the dynamic benchmark.

The paper proceeds as follows. Sections 2–6 examine the four theoretical traditions. Section 7 proves C1. Section 8 derives C2 and C3 and outlines the research programme they generate. Section 9 offers a synthetic assessment. Section 10 maps the open frontiers. Section 11 concludes.

Table 1 provides a compact reference for the four frameworks and the nature of their tensions.

Selection criterion and scope. The selection criterion for inclusion is: a body of work is included if and only if it contributes a distinct mathematical structure to the formal architecture of health economics—state-preference theory, dynamic capital optimisation, constrained social choice, or actuarial intertemporal calculus. This excludes Enthoven (institutional design rather than formal axiomatics), Culyer (a normative competitor to QALY utilitarianism rather than a component of it), and Cutler (empirical estimation of parameters the theoretical frameworks require). The paper is a theoretical synthesis augmented by formal results: it produces no original data. Its empirical claims—that the joint compatibility conditions of Theorem 1 are empirically empty, and that $\delta'(t) > 0$ implies $\lambda^*(t)$ is not constant—follow from regularity assumptions documented in the cited literature.

Table 1: The four frameworks and their policy tensions

Framework	Decision ob- ject	Binding con- straint	Tension with others
Arrow (1963)	Institutional design	First-best unattainable	Diagnoses failure; does not derive threshold or premium
Grossman (1972)	Health investment $I(t)$	Capital accumulation; $\delta(t)$ increasing	Implies age-varying cost profile; undermines flat-premium solvency
Garber–Phelps (1997)	Coverage threshold λ	Fixed budget B	Uniform λ requires income-invariant WTP; ignores demographic dynamics
Zweifel–Pauly (2000, 2009)	Premium Π ; coinsurance α	Actuarial solvency; incentive compatibility	Age-rated premiums conflict with community rating; $\alpha < 1$ generates excess spending under GP

2 Welfare Economics of Medical Care: Uncertainty and Information Asymmetry

2.1 The Canonical Departure from Competitive Equilibrium

Kenneth Arrow's 1963 paper "Uncertainty and the Welfare Economics of Medical Care" remains arguably the most cited contribution to health economics ever published (Arrow, 1963). Its significance lies not merely in identifying empirical deviations from competitive behaviour in medical markets, but in demonstrating that such deviations are structurally necessary consequences of the nature of medical care itself. The paper's central argument is that the medical care market departs from the competitive ideal in ways that are not remediable through standard regulatory interventions, because the deviations arise from features intrinsic to health and illness: their unpredictability, the information asymmetry between physician and patient, and the ethical norms that govern medical practice.

Arrow identifies several distinctive characteristics of medical care that prevent standard welfare-theoretic results from applying. First, the demand for medical care is irregular and unpredictable—it arises from illness, which is itself uncertain in incidence and severity. Second, the commodity supplied is not homogeneous: a unit of medical care is inseparable from the specific physician providing it, creating conditions of non-standardisability that preclude competitive pricing. Third, barriers to entry through licensing, while potentially welfare-enhancing in the presence of the informational problem, constitute a structural departure from the competitive ideal.

Most fundamentally, Arrow identifies information asymmetry as the structural core of the problem. The physician possesses specialised knowledge that the patient lacks; the patient can-

not independently verify whether the diagnosis is correct, whether the recommended treatment is appropriate, or whether an alternative course of action would yield superior outcomes. This informational inequality inverts the usual consumer sovereignty assumption of welfare economics. The patient is, in Arrow's terminology, not a sovereign consumer but a dependent one, whose effective demand is mediated through the judgement of a supplier who is simultaneously agent and expert.

2.2 Insurance and the Response to Uncertainty

Arrow demonstrates that uncertainty about illness creates a demand for insurance, and that competitive insurance markets face systematic difficulties arising from moral hazard and adverse selection. Moral hazard refers to the tendency for insured individuals to consume more medical care than they would in the absence of insurance coverage. Adverse selection refers to the tendency for individuals with higher health risks to disproportionately seek insurance coverage, creating selection effects that undermine the actuarial basis for pooling (Arrow, 1963).

These observations anticipated the foundational contributions of Akerlof (1970) on adverse selection and Rothschild and Stiglitz (1976) on competitive insurance markets under asymmetric information. Arrow's insight that health market failures are systematic and irresolvable by market mechanisms alone provided the intellectual justification for the extensive role of government in health care financing and regulation observed in virtually every advanced economy.

From a mathematical perspective, Arrow's framework draws on the state-preference theory of uncertainty (Arrow, 1953) and the expected utility framework associated with von Neumann and Morgenstern. An individual's utility function is defined over states of the world indexed by health status, and the demand for health insurance can be analysed as a problem of portfolio choice over state-contingent claims. Formally, let $\Omega = \{s_1, \dots, s_n\}$ be a finite state space, $p_i > 0$ the probability of state s_i , and $u_i(c_i)$ the utility derived from consumption c_i in state s_i . The individual maximises expected utility $\sum_i p_i u_i(c_i)$ subject to budget constraints, with insurance providing contingent transfers across states. This formalism underpins subsequent work on optimal insurance contracts, copayments, and benefit package design.

2.3 Arrow's Broader Influence: A Recalibration

Arrow's contribution established the normative benchmark against which health care systems are evaluated. If the competitive market fails to achieve a Pareto-optimal allocation—as Arrow demonstrates it must—then the relevant policy question is: what alternative institutional arrangements can improve upon the market outcome. This framing has structured the health economics literature for six decades.

It is worth noting, as several commentators have observed, that Arrow's 1963 paper anticipated the development of mechanism design theory, principal-agent theory, and contract theory, which found their canonical applications in health economics well before they were formalised

as general theoretical frameworks (Zweifel et al., 2009). The economics of medicine served, in part, as an incubator for general economic theory.

A critical reframing deserves explicit acknowledgement. Arrow saw information asymmetry as a source of *structural failure* requiring institutional intervention. The subsequent contract-theoretic literature—principal-agent theory, mechanism design—reframes the same observation as a *design problem* with engineering solutions: if information is asymmetric, design contracts that align incentives (copayments, deductibles, gatekeeping). These are not compatible research programmes with different vocabularies; they are different diagnoses with different normative implications. Arrow’s framework implies that no contract can achieve first-best efficiency under information asymmetry. Contract theory asks what second-best institutions come closest. The empirical literature on managed care, pay-for-performance, and gatekeeping reflects this unresolved tension: the question is not whether information asymmetry creates problems, but whether those problems admit institutional solutions or merely institutional accommodations.

One limitation of Arrow’s framework that requires acknowledgement: it assumes the physician–patient information asymmetry is durable and irreducible—a structural feature grounded in years of specialised training that patients cannot replicate. This assumption is increasingly contestable in an environment of digital health platforms, AI-assisted diagnosis, and patient-accessible medical records. Whether these technologies erode or entrench the asymmetry—by reducing the *access* gap while leaving the *evaluation* gap intact—is an open question that Arrow’s framework identifies but cannot answer. The structural argument for institutional intervention survives; the specific institutional form it justifies may not.

3 Health Capital Theory: The Demand-Side Foundation

Any synthesis of health economics that omits Grossman’s 1972 health capital model is incomplete. Arrow diagnosed the supply-side failure. Grossman formalised the demand side: why do individuals seek medical care, and how this demand varies across the life cycle.

3.1 Health as a Durable Capital Stock

Grossman (1972) modelled health as a durable capital stock H_t that individuals both inherit and invest in, and that depreciates over time. The individual’s utility function is defined over consumption goods C_t and “healthy time” $\phi(H_t)$ —the fraction of the period available for market and non-market activities conditional on health stock. The optimisation problem is:

$$\max_{\{I_t, C_t\}} \sum_{t=0}^T \beta^t U(\phi(H_t), C_t) \quad (1)$$

subject to the health capital accumulation equation

$$H_{t+1} = I_t + (1 - \delta_t) H_t, \quad (2)$$

where $I_t \geq 0$ denotes gross investment in health (produced using medical care and time), $\delta_t \in (0, 1)$ is the period- t depreciation rate (increasing with age), and β is the subjective discount factor. The stock H_t must remain above a minimum threshold $H_{\min}$ for survival; the individual's endogenous length of life is determined by the first period in which H_t would fall below $H_{\min}$ absent further investment.

3.2 Why Grossman Is Indispensable Here

Three implications of the Grossman model are directly relevant to the synthesis this paper develops.

First, Grossman's framework distinguishes the demand for *health* from the demand for *medical care*. Medical care is one input into health investment; it is not the same thing as health. Arrow's analysis concerns the market for medical care. Grossman explains why that market is demanded in the first place—and therefore what determines the elasticity parameters that Zweifel's optimal coinsurance formula requires.

Second, the depreciation rate δ_t increasing with age is the micro-founded mechanism underlying Zweifel's observation that health care costs rise with age. Zweifel's actuarial models take the age profile of health expenditure as an empirical input; Grossman provides its theoretical foundation. Without Grossman, the actuarial age-cost profile is a statistical regularity without a causal interpretation.

Third, Grossman's framework provides the intertemporal structure that the QALY's discounting formula $Q = \int_0^T u(h(\tau)) e^{-\delta\tau} d\tau$ implicitly assumes. The discount rate δ in the QALY formula is theoretically motivated by the rate of time preference for health—a parameter that Grossman's optimisation characterises explicitly. Garber's cost-effectiveness framework inherits this structure without deriving it; Grossman supplies the derivation.

4 Cost-Effectiveness Analysis and Medical Decision Theory

4.1 From Clinical Uncertainty to Economic Decision Theory

Alan Garber's contributions to health economics extend Arrow's foundational insights by formalising the logic of medical decision-making within a rigorous economic and statistical framework. Garber's work operates at the intersection of clinical medicine, decision theory, and welfare economics, with a particular focus on how the uncertainty inherent in medical care can be rendered tractable for purposes of resource allocation. His central contribution is the formalisation of cost-effectiveness analysis (CEA) as an instrument of health economic evaluation—one that translates clinical outcomes into economically comparable units and provides a basis for rational allocation of health care resources under budget constraints (Garber and Phelps, 1997).

Garber's approach is grounded in the Bayesian tradition of statistical decision theory, which treats the problem of medical decision-making as one of choosing an action from a feasible set in the face of probabilistic uncertainty about outcomes. The physician or planner chooses a treatment strategy that maximises expected utility, where utility is defined over health outcomes and costs. This framework allows the incorporation of uncertainty at multiple levels: uncertainty about the patient's underlying condition, uncertainty about treatment efficacy, and uncertainty about the relationship between clinical outcomes and quality of life (Garber, 2000).

4.2 Cost-Effectiveness Analysis as Economic Evaluation

Cost-effectiveness analysis compares interventions with respect to two dimensions: cost (measured in monetary units) and effectiveness (measured in health outcomes). The incremental cost-effectiveness ratio (ICER) captures the additional cost per unit of health gain associated with moving from one intervention to another:

$$\text{ICER} = \frac{C_1 - C_0}{E_1 - E_0}, \quad (3)$$

where C_1 and C_0 denote the costs of the new and comparator interventions respectively, and E_1 and E_0 denote the corresponding effectiveness measures (typically QALYs). An intervention is deemed cost-effective if $\text{ICER} \leq \lambda$, where λ is the cost-effectiveness threshold.

Garber and Phelps (1997) demonstrate that under specific assumptions about the utility function and the budget constraint, a social planner maximising aggregate health subject to a budget constraint will select all interventions with an ICER below a threshold equal to the value of a statistical life-year. Formally, let the social planner solve

$$\max_{\{x_i\}} \sum_i E_i(x_i) \quad \text{subject to} \quad \sum_i C_i(x_i) \leq B, \quad (4)$$

where x_i is the coverage decision for intervention i , E_i its expected QALY gain, C_i its cost, and B the total budget. The Lagrangian first-order conditions yield $\lambda = \partial B / \partial \sum_i E_i$ —the shadow price of the budget constraint, interpretable as the implicit social value of a QALY. This result provides a theoretical foundation for the cost-effectiveness thresholds used by health technology assessment bodies such as NICE in England and Wales. Importantly, λ is *not* a universal constant; it is determined by the binding budget constraint and the population's preferences, and therefore varies with income, health status, and social preferences—a point to which we return in Section 7.

4.3 Decision Theory, Evidence Synthesis, and Value of Information

Garber's work also engages with the problem of evidence synthesis and the value of information in medical decision-making. In the Bayesian decision-theoretic framework, the value of

additional information is measured by the reduction in expected loss it enables. The expected value of perfect information (EVPI) is:

$$\text{EVPI} = \mathbb{E}_{\theta} \left[\max_d U(d, \theta) \right] - \max_d \mathbb{E}_{\theta} [U(d, \theta)], \quad (5)$$

where θ is the uncertain parameter (e.g., treatment efficacy), d is the decision (e.g., coverage), and U is the utility function. The EVPI measures the gain from learning θ before choosing d , and therefore the maximum a decision-maker should pay for a study that would resolve uncertainty. The expected value of sample information (EVSI) extends this to partial information from a trial of size n (Garber, 2000).

This framework has important implications for research prioritisation. If the EVPI for a proposed clinical trial exceeds its cost, the trial is socially valuable even in the absence of certainty about its outcome. Garber's contributions to this literature have influenced the research prioritisation practices of health technology assessment agencies.

Garber has also contributed to the analysis of disease prevention, examining the conditions under which preventive interventions are cost-effective and the ways in which behavioural factors—including time discounting, risk preferences, and present bias—affect the demand for preventive care (Garber, 2004). This work connects health economics with behavioural economics, extending the standard rational-choice framework to accommodate empirically documented departures from expected utility maximisation—departures whose implications for the QALY are discussed directly in Section 5.4.

5 QALY and WTP: Measuring Health Outcomes in Economic Terms

5.1 The Quality-Adjusted Life Year

The Quality-Adjusted Life Year (QALY) is the dominant measure of health outcomes in economic evaluation, and its theoretical foundations are deeply entwined with the contributions of both Arrow and Garber. A QALY combines information about the duration and quality of life into a single composite measure. For a health trajectory $h(\tau)$ over a lifetime horizon T :

$$Q = \int_0^T u(h(\tau)) e^{-\delta\tau} d\tau, \quad (6)$$

where $u(h) \in [0, 1]$ is the utility weight assigned to health state h , with $u = 1$ representing full health and $u = 0$ representing a state equivalent to death, and δ is the discount rate applied to future health benefits (Drummond et al., 2015).

The welfare-theoretic foundations of the QALY trace back to the expected utility framework of von Neumann and Morgenstern (1944), which provides the mathematical grammar without

which the QALY cannot be stated as a welfare measure. The QALY is a representation of a von Neumann–Morgenstern utility function defined jointly over survival duration and health quality; without expected utility theory, the numerical weights $u(h)$ are merely ratings, not welfare indices. Pliskin et al. (1980) completed the foundational work by establishing the *if and only if* conditions for this representation: the QALY model corresponds to a von Neumann–Morgenstern utility function over duration and health status if and only if three conditions hold simultaneously: (i) *utility independence*: life duration is utility-independent of health status; (ii) *constant proportional risk-posture*: the certainty equivalent of any risky prospect over life duration is a constant proportion of the expected duration; and (iii) *risk neutrality* over quality adjustments. These are strong conditions—not the weakest sufficient conditions one might hope for, but necessary ones. Their potential violation does not merely limit the QALY; it removes its welfare-theoretic interpretation entirely, reducing it to an index number without normative content. Much of the applied literature has proceeded pragmatically alongside these restrictive representation conditions, prioritising tractability over foundational consistency.

5.2 Methods of Health State Valuation

The utility weights assigned to health states can be elicited through several methodological approaches, each with distinct theoretical properties. The Time Trade-Off (TTO) method asks respondents to indicate the number of years in full health they regard as equivalent to a longer period in a given health state: if an individual is indifferent between t^* years in full health and T years in health state h , then $u(h) = t^*/T$. The Standard Gamble (SG) method, which is more directly grounded in expected utility theory, asks respondents to indicate the probability p of full health—against a complementary probability of immediate death—such that they are indifferent between the gamble and the certainty of health state h . The resulting utility weight is $u(h) = p$ (Torrance, 1986).

The Visual Analogue Scale (VAS) method, which elicits cardinal ratings on a scale from 0 to 100, lacks direct expected utility foundations and its weights are generally not suitable for use in expected value calculations without further transformation. Empirically, TTO and SG methods tend to yield systematically different utility weights, with SG weights typically exceeding TTO weights—a finding that may reflect the differential role of risk aversion in the two elicitation procedures, as well as the presence of loss aversion (discussed below).

5.3 Willingness to Pay and the Value of a Statistical Life

An alternative approach to valuing health outcomes is the Willingness to Pay (WTP) method, grounded in the compensating variation concept from welfare economics. For a change in health from h_0 to h_1 , the WTP measure W satisfies:

$$V(h_1, y - W) = V(h_0, y), \quad (7)$$

where $V(h, y)$ is the indirect utility function defined over health status h and income y , and W is the compensating variation associated with the health improvement (Johannesson, 1996).

Garber and Phelps (1997) establish conditions under which a cost-utility threshold derived from WTP considerations corresponds to the marginal value of a QALY. The key insight is that the value of a QALY—the threshold λ in Section 4—is not a universal constant but depends on the individual's income, health status, and preferences. This result has a distributional implication that the subsequent literature has not fully confronted: if richer individuals have higher WTP for a QALY, a uniform threshold implicitly overvalues health improvements accruing to the wealthy and undervalues those accruing to the poor, or vice versa depending on which direction the calibration errs. The WTP and uniform-QALY frameworks are not reconcilable without specifying a social welfare function that makes a distributional judgement explicit (see Section 7.1).

The Value of a Statistical Life (VSL) provides a population-level aggregation of individual WTP estimates for mortality risk reductions. If each individual in a population of N persons is willing to pay WTP_i for a reduction of $1/N$ in the risk of death, the aggregate WTP equals the $VSL = WTP_i$. The VSL is the primary basis for the monetisation of mortality risk reductions in regulatory impact analyses (Viscusi and Aldy, 2003).

The calibration of VSL estimates to QALY-based thresholds is not a technical detail; it is a foundational problem for the coherence of health economic evaluation. The difficulty can be stated precisely. Let λ be the cost-effectiveness threshold and μ the expected remaining QALYs for an individual facing a small mortality risk reduction $\Delta\pi$. A naive calibration sets:

$$VSL = \lambda \cdot \mu, \quad (8)$$

implying that the monetary value of avoiding a statistical death equals the threshold times the expected QALY loss from that death. Equation (8) is internally consistent only under three joint restrictions. First, the utility function must be additively separable in health quality and survival duration, so that the QALY formula (6) represents the relevant welfare object. Second, the marginal utility of income must be constant, so that individual WTP for a mortality risk reduction is income-invariant and the VSL does not vary across the income distribution. Third, the discount rate δ in the QALY formula must equal the rate implicit in hedonic wage-risk estimates of the VSL.

None of these restrictions holds in practice. The first is the Pliskin separability condition that behavioural evidence systematically undermines. The second contradicts the Garber–Phelps finding that $\lambda(y, h)$ is strictly increasing in income. The third is empirically unsettled: hedonic wage studies reflect workers' revealed discount rates for mortality risk, which need not equal the 3.5 or 5 per cent rates conventionally applied in cost-effectiveness calculations (Viscusi and Aldy, 2003). The implication is that any jurisdiction using both a cost-effectiveness threshold for coverage decisions and a VSL for regulatory impact assessment is simultaneously applying

two incommensurable monetary valuations of human life. Resolving this requires specifying a single welfare function that generates both objects—a programme the literature has not completed.

5.4 The Behavioural Challenge to QALY Foundations

The challenge from behavioural economics is not a future complication to be deferred; it is a present problem for the QALY's welfare-theoretic foundations, active at the moment utility weights are elicited from respondents.

Three empirically documented departures from expected utility theory are directly relevant. *Loss aversion*: individuals weight health losses more heavily than equivalent health gains—a finding that violates the symmetry between TTO and SG that the QALY framework requires. *Probability weighting*: individuals overweight small probabilities and underweight large ones, so that the SG utility weight $u(h) = p$ is not a linear function of the objective probability, as Pliskin et al.'s derivation requires. *Present bias*: individuals discount near-term health outcomes more steeply than the constant δ in equation (6) implies, generating time-inconsistent preferences that the stationary QALY model cannot accommodate.

The standard response in the health economics literature has been to acknowledge these violations in footnotes while continuing to use the framework. Two principled alternatives exist, and neither has been fully developed.

The first response is *normative defence*: accept that individuals violate expected utility theory descriptively, but maintain that the normative standard for resource allocation should nevertheless be expected utility, corrected for “decision errors.” This is a defensible position—paternalistic welfare economics has a long tradition—but it requires an explicit argument for why the planner's preferences should override the patient's revealed preferences, particularly for the TTO and SG weights that are directly elicited from affected individuals.

The second response is *behavioural reconstruction*: replace the von Neumann–Morgenstern utility function with a rank-dependent or prospect-theoretic representation and re-derive the conditions under which a QALY-like measure exists. This programme has not been completed. It is not obvious that a “behavioural QALY” would preserve the additivity property $Q = \int_0^T u(h(\tau)) e^{-\delta\tau} d\tau$ that makes QALYs computationally tractable and institutionally usable. If additivity fails, the entire apparatus of ICER comparison and budget-constrained optimisation would require reconstruction.

The uncomfortable conclusion is that the QALY is institutionally stable and theoretically unstable simultaneously. Institutional practice has prioritised tractability and cross-study comparability over full consistency with behavioural decision theory—a defensible choice, but one that should be stated explicitly rather than left implicit.

5.5 Equity Considerations in QALY-Based Evaluation

Standard cost-effectiveness analysis maximises aggregate QALYs subject to a budget constraint, treating each QALY as equally valuable regardless of the characteristics of the recipient. This embodies an implicit utilitarian social welfare function of the form $W = \sum_i Q_i$, which many analysts regard as insufficiently sensitive to distributional concerns.

Alternatives have been proposed, including distributional weights that assign greater value to QALYs accruing to individuals with lower initial health endowments, younger ages, or lower socioeconomic status—analogous to the Atkinson family of social welfare functions used in the analysis of income distribution. NICE has incorporated elements of severity weighting, granting greater weight to QALYs generated for patients with severe conditions (Claxton et al., 2015). This represents a partial accommodation of non-utilitarian distributional preferences within the ICER framework. As noted in Section 7, however, non-additive distributional weights are incompatible with simple ICER-based ranking: once weights differ across patients, maximising weighted QALYs cannot be reduced to a threshold comparison.

6 The Actuarial Foundations of Health Insurance

6.1 Mathematical Economics of Insurance and Risk

Peter Zweifel's contributions to health economics are distinguished by their mathematical rigour and systematic integration of actuarial science with microeconomic theory (Zweifel et al., 2009). The fundamental problem of health insurance can be formulated as follows. Let an individual face a random health expenditure X , drawn from a distribution $F(x)$ with mean μ and variance σ^2 . In the absence of insurance, the individual bears the full cost of illness. Under a complete insurance contract, the individual pays a premium $P = \mu$ in exchange for full coverage. The welfare gain from insurance is measured by the difference between the certainty equivalent of the individual's lottery over health expenditures and the actuarially fair premium—a gain positive for risk-averse individuals, increasing with the degree of risk aversion and the variance σ^2 .

Zweifel formalises this within the Arrow–Debreu framework of state-contingent markets, establishing conditions under which competitive insurance markets yield Pareto-optimal allocations. Complete insurance markets—in which a contingent claim can be traded for every possible health state—yield a Pareto-optimal allocation. In practice, the incompleteness of markets, the presence of moral hazard, and information asymmetries between insurer and insured prevent this ideal from being achieved.

6.2 Moral Hazard, Adverse Selection, and Optimal Contract Design

Zweifel's analysis of moral hazard builds on Pauly (1968) and provides a rigorous treatment of the efficiency costs of insurance-induced overconsumption of medical care. Under full insurance, the marginal cost of health care to the consumer is zero while the marginal social cost is positive—a wedge that produces excessive consumption. The optimal coinsurance rate α^* balances risk-sharing benefits against moral hazard costs, and is determined by the elasticity of demand for medical care with respect to price:

$$\alpha^* = \frac{\varepsilon_d / |\varepsilon_d| \cdot \rho}{1 + \varepsilon_d / |\varepsilon_d| \cdot \rho}, \quad (9)$$

where $\varepsilon_d < 0$ is the price elasticity of demand for medical care and ρ captures the degree of risk aversion. In the principal–agent formulation, the insurer (principal) designs a contract (Π, α) specifying a premium Π and a coinsurance rate α , maximising expected profit subject to the individual rationality constraint (the insured prefers the contract to no insurance) and the incentive compatibility constraint (the insured cannot profitably misrepresent their type). The first-order condition equates the marginal benefit of additional risk protection with the marginal cost of additional moral hazard distortion (Zweifel and Manning, 2000).

Two challenges to the standard moral hazard framework require explicit acknowledgement.

The first is the *income-effect critique*, associated with Nyman (2001), which argues that Pauly's welfare analysis misclassifies a portion of insurance-induced consumption as inefficient. When an individual falls ill, insurance acts as an income transfer: it moves resources from the healthy state (the premium) to the sick state (the benefit), enabling the sick individual to purchase more medical care than they could afford uninsured. This income effect is not a distortion; it is the mechanism through which insurance achieves its welfare improvement. Pauly's framework attributes all insurance-induced consumption to the price distortion (moral hazard), but the income and price effects are, in principle, observationally equivalent. The optimal coinsurance rate formula α^* is therefore calibrated against an efficiency concept whose welfare decomposition is unsettled. No consensus exists on how to decompose the two effects empirically.

The second challenge concerns the empirical inputs the theoretical framework requires. The optimal coinsurance formula is a function of the price elasticity of demand for medical care ε_d —a parameter that Zweifel's model treats as known. The primary source of credible estimates of ε_d is the RAND Health Insurance Experiment, a large-scale randomised trial conducted in the United States in the 1970s and 1980s, which is not cited in this paper but is the empirical foundation on which Zweifel's theoretical results rest. Without RAND, the optimal coinsurance formula is a function of an unknown parameter. The formula is theoretically clean; the empirical input it requires is poorly identified outside the RAND context, varies substantially across populations and health conditions, and reflects a study population from four decades ago.

The current institutional frontier—value-based insurance design (VBID)—accepts the Pauly–Zweifel framework but applies it selectively: reducing cost-sharing for high-value services (those with low ICERs or high elasticity-adjusted welfare gains) and increasing it for low-value services. VBID does not resolve the income-effect critique—it inherits Pauly’s welfare decomposition—but it represents a practical accommodation of the framework’s insight that not all coinsurance rates should be uniform.

Adverse selection in health insurance arises when individuals possess private information about their health risks that is unobservable to insurers. Zweifel’s treatment draws on Rothschild and Stiglitz (1976) and extends it to accommodate the multi-dimensional nature of health risk, the role of age and chronic conditions in risk stratification, and the implications of community rating regulations. The key theoretical result is that competitive equilibrium in the presence of adverse selection either does not exist or involves inefficient separating equilibria in which low-risk individuals accept incomplete coverage to signal their type (Zweifel et al., 2009).

6.3 Long-Run Sustainability and Actuarial Solvency

Zweifel’s most distinctive contribution concerns the long-run actuarial sustainability of insurance systems. Pay-as-you-go systems face systematic fiscal pressures from demographic ageing, technological progress, and the secular growth of health care costs. The actuarial imbalance of a health insurance system is:

$$AI = PV(\text{projected expenditures}) - PV(\text{projected contributions}), \quad (10)$$

where both present values are computed over an infinite horizon at a risk-adjusted discount rate. A positive actuarial imbalance signals long-run insolvency.

Zweifel develops a formal intertemporal model of health insurance finance in which the premium path is determined by the requirement that the expected present value of lifetime contributions equals the expected present value of lifetime benefits for each cohort:

$$\mathbb{E} \left[\int_0^{\bar{T}} \Pi(t) e^{-rt} dt \right] = \mathbb{E} \left[\int_0^{\bar{T}} B(t) e^{-rt} dt \right], \quad (11)$$

where $\Pi(t)$ is the premium at age t , $B(t)$ is the expected benefit at age t , r is the discount rate, and $\bar{T}$ is the expected length of life. This intergenerational equity condition implies that premiums must be age-rated to reflect the increasing health care costs associated with ageing—a requirement that conflicts with community rating regulations prevalent in many countries. The tension between actuarial fairness and political equity is a central theme in Zweifel’s analysis (Zweifel et al., 2009).

7 Tensions Between the Three Frameworks

The preceding sections have presented three traditions as complementary. They are not. This section formalises the tensions between them.

7.1 The WTP–QALY Incompatibility

Garber and Phelps (1997) demonstrate that the value of a QALY—the threshold λ in equation (3)—is not a universal constant but a function of income y and health status h :

$$\lambda(y, h) = \frac{\partial V / \partial h}{\partial V / \partial y} \cdot \frac{1}{\Delta Q}, \quad (12)$$

where $V(h, y)$ is the indirect utility function and ΔQ is the QALY gain from the intervention. Under standard assumptions about the marginal utility of income ($\partial^2 V / \partial y^2 < 0$), λ is strictly increasing in income: richer individuals implicitly value a QALY more.

A uniform threshold λ^* used in practice (e.g., NICE’s £20–30k range) is therefore theoretically indefensible as a welfare measure: it undervalues health gains to individuals with $\lambda(y, h) > \lambda^*$ and overvalues those with $\lambda(y, h) < \lambda^*$. The only resolutions are either to use income-differentiated thresholds—which requires a distributional judgement that technocratic agencies prefer to avoid, and implies that wealthier patients’ lives are institutionally worth more—or to adopt a social welfare function with explicit redistributive weights that decouples the social value of a QALY from individual WTP. The QALY framework and the WTP framework are not reconcilable without such a choice.

7.2 The Actuarial Fairness–Community Rating Conflict

Zweifel’s solvency condition (11) implies that premiums must be age-rated. Community rating—the regulation prohibiting age-differentiated premiums, operative in various forms in most European health systems—directly violates this condition. Within Zweifel’s own model, a community-rated system cannot satisfy equation (11) unless young cohorts are taxed to subsidise older cohorts, and the size of this implicit intergenerational transfer grows with the degree of demographic ageing. Zweifel identifies this as a “central tension”; it is more precisely a logical incompatibility within the same formal framework.

7.3 Formal Setup

We introduce notation that will carry through the formal results. Let $V : \mathcal{H} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ be an indirect utility function defined over health status $h \in \mathcal{H} \subseteq [0, 1]$ and income $y \in \mathbb{R}_+$, twice continuously differentiable. Let $B(t)$ denote expected health expenditure for an individual of age $t \in [0, \bar{T}]$, derived from Grossman’s health capital model with depreciation $\delta(t)$. Let $Q_j(P_c)$

denote aggregate demand for medical service j at consumer price P_c , with social marginal cost P_j . Let $B > 0$ denote the social planner's health care budget. We impose three regularity assumptions:

- (A1) $V_{yy}(h, y) < 0$ for all (h, y) : diminishing marginal utility of income.
- (A2) $\delta'(t) > 0$ for all $t \in [0, \bar{T}]$: health capital depreciates at a strictly increasing rate with age, so that $B'(t) > 0$.
- (A3) $\varepsilon_d(P_c) \equiv \frac{\partial Q_j}{\partial P_c} \frac{P_c}{Q_j} < 0$ for all $P_c > 0$ and all services j : demand for medical care is price-elastic.

Assumptions (A1)–(A3) are not strong; they are the weakest conditions consistent with the empirical record. (A1) is a standard assumption in welfare economics. (A2) is the micro-economic implication of Grossman's health capital model for the observed positive relationship between age and health care costs. (A3) is supported by the RAND Health Insurance Experiment and the subsequent quasi-experimental literature.

We now state three framework conditions formally.

Definition 1 (GP-Optimality). A health system is *Garber–Phelps optimal* if the social planner implements a universal threshold $\lambda^* \in \mathbb{R}_+$ such that:

- (i) service j is covered if and only if $\text{ICER}_j \leq \lambda^*$; and
- (ii) the budget constraint binds: $\sum_{j: \text{ICER}_j \leq \lambda^*} P_j \cdot Q_j(P_j) = B$.

Here λ^* is the shadow price of the budget constraint, equal to $(\partial V / \partial h) / (\partial V / \partial y) \cdot (1 / \Delta Q)$ evaluated at the social planner's optimum. This is the *WTP-based* interpretation of the threshold, in which the social value of a QALY is derived from individuals' marginal rates of substitution between health and income, without additional redistributive weights. Results derived from this definition do not apply to threshold concepts based on opportunity cost (Claxton et al., 2015) or extra-welfarist social welfare functions.

Definition 2 (ZS-Solvency). An insurance system with community rating ($\Pi(t) = \bar{\Pi}$ for all t) is *Zweifel-solvent with intergenerational equity* if, for each birth cohort c :

$$\mathbb{E} \left[\int_0^{\bar{T}} \bar{\Pi} e^{-rt} dt \mid \text{cohort } c \right] = \mathbb{E} \left[\int_0^{\bar{T}} B(t) e^{-rt} dt \mid \text{cohort } c \right]. \quad (13)$$

Definition 3 (PZ-Incentive Compatibility). A coinsurance rate $\alpha^* \in (0, 1)$ is *Pauly–Zweifel incentive compatible* if it equates the marginal benefit of additional risk protection with the marginal cost of additional moral hazard distortion—i.e., it satisfies the first-order condition:

$$\frac{\partial}{\partial \alpha} [\text{Risk premium}(\alpha) - \text{DWL}(\alpha)] \Big|_{\alpha=\alpha^*} = 0, \quad (14)$$

where $DWL(\alpha) = \frac{1}{2}(1 - \alpha)^2 P_j^2 |\varepsilon_d| Q_j(P_j)/P_j$ is the Harberger dead-weight loss from demand distortion, and $Risk\ premium(\alpha)$ is the actuarially valued benefit of partial insurance relative to no insurance. Under (A3), the unique solution satisfies $\alpha^* \in (0, 1)$.

7.4 The Joint Incompatibility Theorem [*Contribution C1*]

Theorem 1 (Joint Incompatibility). *Under assumptions (A1)–(A3), no health system simultaneously satisfies GP-Optimality (Definition 1), ZS-Solvency (Definition 2), and PZ-Incentive Compatibility (Definition 3).*

We establish Theorem 1 through three lemmas, each showing that one framework condition is violated when the other two hold under (A1)–(A3).

Lemma 1 (GP requires income-invariant MRS). *Under (A1), the universal threshold λ^* required by Definition 1 cannot exist.*

Proof. From Definition 1, λ^* is universal—income-invariant—only if $\partial\lambda(y, h)/\partial y = 0$ for all (y, h) . Differentiating:

$$\frac{\partial\lambda}{\partial y} = \frac{1}{\Delta Q} \cdot \frac{V_{yh} V_y - V_h V_{yy}}{V_y^2}. \quad (15)$$

Setting $\partial\lambda/\partial y = 0$ requires $V_{yh} V_y = V_h V_{yy}$, or equivalently:

$$\frac{V_{yh}}{V_h} = \frac{V_{yy}}{V_y}. \quad (16)$$

Under (A1), $V_{yy} < 0$ and $V_y > 0$, so the right-hand side of (16) is strictly negative. Since $V_h > 0$ (health improves utility), (16) requires $V_{yh} < 0$. However, $V_{yh} < 0$ implies that the marginal utility of health is *decreasing* in income—richer individuals value an additional unit of health less. This is the opposite of the empirical finding of Garber and Phelps (1997), who demonstrate that $\lambda(y, h)$ is strictly increasing in income, requiring $V_{yh} > 0$. The condition $V_{yh} < 0$ is therefore contradicted by the empirical evidence that motivates Definition 1 in the first place. Hence $\partial\lambda/\partial y > 0$ under any empirically supported V , and a universal λ^* cannot be derived from WTP-based aggregation as in Definition 1.

Two clarifications are warranted. First, the result applies to the WTP interpretation of Definition 1, not to every possible justification for a uniform threshold. A uniform threshold could be defended on opportunity-cost grounds (Claxton et al., 2015), as a second-best approximation under administrative constraints, or via a social welfare function with explicit equity weights that decouples the threshold from individual income. Second, the result does not imply that NICE or CADTH act irrationally; it implies that their uniform thresholds require a normative foundation that is not supplied by the WTP-based welfare calculus of Garber and Phelps. $\square$

Lemma 2 (ZS-Solvency under CR requires flat age-cost profile). *Under (A2), ZS-Solvency (Definition 2) under community rating cannot hold with intergenerational equity.*

Proof. For each birth cohort to be actuarially self-financing (equation (13)), the community premium $\bar{\Pi}$ must satisfy, for every cohort:

$$\bar{\Pi} \cdot \mathbb{E} \left[\int_0^{\bar{T}} e^{-rt} dt \right] = \mathbb{E} \left[\int_0^{\bar{T}} B(t) e^{-rt} dt \right]. \quad (17)$$

Since $\bar{\Pi}$ is the same for all cohorts (community rating) but the right-hand side of (17) varies with the age distribution of the cohort, equation (17) can hold for all cohorts simultaneously only if $B(t)$ is constant in t —so that every cohort's expected present-value benefits equal $\bar{\Pi}$ times the same expected discount factor. Formally, let $a = \mathbb{E}[\int_0^{\bar{T}} e^{-rt} dt]$. Then (17) gives $\bar{\Pi} = \mathbb{E}[\int_0^{\bar{T}} B(t) e^{-rt} dt]/a$ for every cohort, which requires $\int_0^{\bar{T}} B(t) e^{-rt} dt$ to be the same for all cohorts—possible only if $B(t)$ is constant. Under (A2), however, $\delta'(t) > 0$ implies that the gross health investment required to maintain H_t above the minimum threshold $H_{\min}$ strictly increases with age, so $B'(t) > 0$ (Grossman, 1972). Contradiction. $\square$

Lemma 3 (PZ coinsurance violates the GP budget constraint under unconstrained utilisation). *Under (A3), PZ-Incentive Compatibility (Definition 3) implies $\alpha^* \in (0, 1)$. If utilisation is unconstrained by quantity controls, prior authorisation, or physician gatekeeping, then aggregate spending on covered services exceeds the GP budget B , violating Definition 1(ii). Real health systems employ such mechanisms alongside coinsurance; the lemma establishes that the Pauly–Zweifel coinsurance rule and the Garber–Phelps budget allocation rule cannot be simultaneously implemented without an additional utilisation-control instrument.*

Proof. Under (A3) and Definition 3, the first-order condition (14) admits a unique interior solution $\alpha^* \in (0, 1)$: full insurance ($\alpha = 0$) is never optimal because the moral hazard cost $\text{DWL}(0) > 0$ under (A3), and no insurance ($\alpha = 1$) forfeits all risk-sharing benefit.

A preliminary remark on the role of B . In both GP-Optimality and PZ-Incentive Compatibility, B is the *same exogenously fixed* social budget—the total resources society has committed to health care. It is not a free parameter that adjusts to accommodate whichever coinsurance rate the insurer sets. GP takes B as given and derives the coverage set; PZ takes B as given and derives α^* . The incompatibility arises because these two operations, applied to the same B , yield contradictory demands on utilisation. The objection that B could be redefined to equal $S(\alpha^*)$ would dissolve the GP framework entirely: it would mean the planner covers whatever consumers choose to consume at α^* , abandoning the budget-constrained optimisation that defines GP-Optimality.

Under $\alpha^* \in (0, 1)$, consumers face a consumer price $\alpha^* P_j < P_j$ for each covered service j . By (A3):

$$Q_j(\alpha^* P_j) > Q_j(P_j) \quad \text{for all covered services } j. \quad (18)$$

Total social spending on covered services under PZ coinsurance is therefore:

$$S(\alpha^*) = \sum_{j: \text{ICER}_j \leq \lambda^*} P_j \cdot Q_j(\alpha^* P_j) > \sum_{j: \text{ICER}_j \leq \lambda^*} P_j \cdot Q_j(P_j) = B, \quad (19)$$

where the equality follows from Definition 1(ii), which pins total spending at B under social-cost pricing. Since B is exogenous and the same in both frameworks, $S(\alpha^*) > B$ constitutes a genuine violation of the GP budget constraint, not a labelling artefact. $\square$

Proof of Theorem 1. By Lemma 1, under (A1) no universal threshold λ^* exists under any empirically supported indirect utility function, so GP-Optimality fails. By Lemma 2, under (A2) ZS-Solvency with intergenerational equity under community rating fails because $B'(t) > 0$. By Lemma 3, under (A3) PZ-Incentive Compatibility induces $\alpha^* \in (0, 1)$, which generates excess spending $S(\alpha^*) > B$, violating the GP budget constraint. Each lemma derives a contradiction from one of (A1)–(A3) independently; together, they establish that no health system can simultaneously satisfy all three framework conditions under any combination of (A1)–(A3). $\square$ $\square$

Corollary 1 (Empirical Emptiness of the Joint Compatibility Condition). *The set of indirect utility functions, age-cost profiles, and demand functions that simultaneously satisfy GP-Optimality, ZS-Solvency, and PZ-IC is empirically implausible. The joint compatibility condition requires $V_{yh} < 0$, $B'(t) = 0$, and $\varepsilon_d = 0$, all of which are contradicted by the empirical evidence cited in this paper. The qualifier empirically implausible rather than logically impossible reflects that each condition fails under the standard empirical evidence; alternative evidence could, in principle, reverse individual conditions, though their joint satisfaction would require a simultaneous coincidence of effects that has not been documented in any health system.*

Remark 1 (What the theorem does not claim). Four things Theorem 1 does not assert deserve explicit statement. (i) It does not imply that mathematical health economics is incoherent. The three frameworks address genuinely different decision problems that are each internally rigorous. (ii) It does not imply that existing health systems are irrational. Real systems resolve the tensions through additional instruments—distributional weights, quantity controls, political compromises—that lie outside each individual framework. (iii) It does not apply to threshold concepts other than the WTP-based Garber–Phelps definition (see Definition 1 note); opportunity-cost thresholds (Claxton et al., 2015) and need-based allocations are not covered. (iv) It does not establish that no unified framework can exist; it establishes that the three benchmark rules cannot, taken together, constitute such a framework. The positive programme of Section 8 addresses what a unified framework would require.

7.5 The Utilitarian–Distributional Conflict in QALY Aggregation

The NICE severity weighting system illustrates a further incompatibility. Standard CEA maximises $\sum_i Q_i$ —an additive, utilitarian objective. Severity weighting assigns weight $w_i > 1$ to

QALYs for severely ill patients, so the objective becomes $\sum_i w_i Q_i$. But once weights differ across patients, ICER-based allocation is no longer equivalent to maximising the weighted sum: two interventions with the same ICER may yield different values of $\sum_i w_i Q_i$ depending on the severity distribution of beneficiaries. The institutional adoption of severity weighting therefore requires either abandoning ICER-based threshold comparison or restricting the weight function to cases where ICER comparison remains consistent with weighted maximisation. This restriction is not generally satisfied (Claxton et al., 2015).

8 Toward a Dynamic Theory of Cost-Effectiveness: A Research Programme

Theorem 1 establishes what the three frameworks, taken together, *cannot* do. This section asks what a framework that could do it would look like — specifically, what structure is required for the cost-effectiveness threshold to be determined endogenously rather than set exogenously by a budget committee.

8.1 The Unified Dynamic Problem

Let the social planner maximise discounted social welfare over an infinite horizon:

$$\max_{\{I(t), \alpha(t), \mathcal{C}(t)\}} \int_0^\infty e^{-\rho t} W(H(t), y(t)) dt, \quad (20)$$

subject to four joint constraints. First, the Grossman health capital dynamics:

$$\dot{H}(t) = I(t) - \delta(t)H(t), \quad H(0) = H_0, \quad (21)$$

where $I(t) \geq 0$ is gross health investment and $\delta(t) > 0$ is the age-specific depreciation rate (Assumption A2). Second, the instantaneous budget constraint:

$$\sum_{j \in \mathcal{C}(t)} P_j Q_j (\alpha(t) P_j) \leq B(t), \quad (22)$$

where $\mathcal{C}(t) = \{j : \text{ICER}_j(t) \leq \lambda(t)\}$ is the set of covered interventions. Third, Zweifel's actuarial solvency condition (13). Fourth, Pauly–Zweifel incentive compatibility for the coinsurance rate $\alpha(t) \in (0, 1)$.

We impose one additional technical assumption that connects health expenditure to health capital:

- (T) *Unit production of health investment*: one unit of health care expenditure translates into one unit of gross health investment, so $I(t) = C(t)$, where $C(t)$ is total health care expenditure. This is a simplification; a richer model would specify a differentiable production

function $I(t) = f(C(t))$ with $f' > 0$, yielding the first-order condition $\mu_H(t) f'(C^*(t)) = \lambda_B(t)$ at the optimum. Under (T), $f' \equiv 1$, so $\mu_H(t) = \lambda_B(t)$.

The key object is $\lambda(t) \equiv \lambda_B(t)$, the Lagrange multiplier on constraint (22). In the static Garber–Phelps framework, λ^* is determined once and for all by the level of the budget B . In the dynamic problem (20)–(22) under (T), $\lambda(t)$ is an endogenous co-state variable whose path is governed by the joint dynamics of health capital and budget.

8.2 The Endogenous Threshold [Contribution C2]

Proposition 1 (Dynamic Cost-Effectiveness Threshold). *In the unified dynamic programme (20)–(22), the optimal cost-effectiveness threshold $\lambda^*(t)$ satisfies the forward-looking integral:*

$$\lambda^*(t) = \int_t^\infty W_H(s) \exp\left(-\int_t^s [\rho + \delta(u)] du\right) ds, \quad (23)$$

where $W_H(s) = \partial W / \partial H(s)$ is the marginal social value of health at time s , and $\rho + \delta(u)$ plays the role of the discount rate for health investments at age u .

Proof sketch. Form the current-value Hamiltonian with co-state variable μ_H for health capital and Lagrange multiplier $\lambda_B(t)$ for the budget constraint:

$$\mathcal{H}_c = W(H, y) + \mu_H [C(t) - \delta(t)H] - \lambda_B(t) \left[\sum_{j \in \mathcal{C}} P_j Q_j (\alpha P_j) - B \right],$$

where we used $I(t) = C(t)$ from Assumption (T). The first-order condition for $C(t)$ gives

$$\frac{\partial \mathcal{H}_c}{\partial C} = \mu_H(t) - \lambda_B(t) = 0 \implies \mu_H(t) = \lambda_B(t) \equiv \lambda^*(t).$$

The co-state equation for health capital is $\dot{\mu}_H = \rho \mu_H - W_H + \mu_H \delta(t)$, which under $\mu_H = \lambda^*$ becomes

$$\dot{\lambda}^*(t) = \lambda^*(t) [\rho + \delta(t)] - W_H(t).$$

This is a first-order linear ODE in λ^* . Solving forward with the transversality condition

$$\lim_{t \rightarrow \infty} e^{-\rho t} \lambda^*(t) H(t) = 0$$

gives equation (23). □ □

Corollary 2 (Demographic Dependence of the Threshold — Contribution C3). *Under Assumption (A2) ($\delta'(t) > 0$), with the marginal social value of health $W_H(s)$ held fixed along the demographic trajectory, the dynamic threshold $\lambda^*(t)$ is strictly decreasing in the depreciation rate $\delta(\cdot)$: an upward shift in the age-specific depreciation profile lowers $\lambda^*(t)$ for all t through the*

depreciation channel of equation (23). As populations age and $\delta(t)$ rises, the effective discount rate for health investments increases, reducing the present value of future health benefits and thus the maximum justifiable expenditure per QALY.

Qualification: the total effect of demographic ageing on $\lambda^*(t)$ may be offset if $W_H(s)$ rises sufficiently with age—for instance, if social welfare weights assign greater value to health of older individuals, or if proximity to death raises the marginal value of health investment. The corollary establishes the sign of the depreciation channel unconditionally; the net effect requires further assumptions on $W_H(\cdot)$. Uniform thresholds are dynamically suboptimal unless both $W_H(t)$ and $\delta(t)$ happen to be constant—a case excluded by (A2) and standard evidence on age-varying health valuations.

Three implications follow directly.

First, $\lambda^*(t)$ is not a constant. The standard HTA practice of applying a time-invariant threshold (NICE's £20–30k, CADTH's CA\$50k) is dynamically inconsistent with the unified optimisation problem: it can coincide with $\lambda^*(t)$ only at a single instant, not along a demographic trajectory.

Second, the incompatibility of Theorem 1 can now be located precisely within the dynamic framework. The Garber–Phelps condition requires $\lambda^*(t)$ to be uniform across individuals (Lemma 1); Zweifel's solvency condition requires premiums to track $B(t)$ over the demographic trajectory; and Pauly–Zweifel coinsurance requires $\alpha^*(t)$ to vary with the demand elasticity. All three are simultaneously satisfiable in the dynamic problem only if W_H , $\delta(t)$, and ε_d satisfy a joint condition that Corollary 1 shows is empirically empty. The dynamic formulation does not resolve the incompatibility; it reexpresses it as an explicit constraint on the path of $\lambda^*(t)$, which is itself the variable that policy can, in principle, choose.

Third, the endogenous threshold connects to Grossman's framework in a way the static models obscure. Because $\delta(t)$ is the depreciation rate of health capital and ρ is the social discount rate, the denominator $\rho + \delta(t)$ in the integrand of (23) is precisely the *user cost of health capital* in Grossman's sense. The optimal cost-effectiveness threshold is therefore the present value of future marginal health benefits discounted at the user cost of health capital—a result that integrates the allocation and investment dimensions of health economics into a single object.

8.3 The Dynamic ICER

The static ICER measures the cost per QALY of an intervention evaluated at a single point in time. In the dynamic framework, the effectiveness of intervention j at time t must account for the entire future health capital trajectory it induces. Define the dynamic effectiveness measure as:

$$E_j^{\text{dyn}}(t) = \int_t^\infty [u(H_j(s)) - u(H_0(s))] \exp\left(-\int_t^s [\rho + \delta(u)] du\right) ds, \quad (24)$$

where $H_j(s)$ is the health capital trajectory under intervention j and $H_0(s)$ is the baseline. The dynamic ICER is then $\text{ICER}_j^{\text{dyn}}(t) = C_j/E_j^{\text{dyn}}(t)$.

The comparison with the static ICER reveals a systematic bias. Static analysis uses a constant discount rate δ ; dynamic analysis discounts at $\rho + \delta(u)$, which increases with age. Static ICER therefore *overstates* the effectiveness of interventions in ageing populations relative to the dynamic benchmark, because it does not account for the accelerating depreciation of the health capital the intervention creates. This provides a theoretical foundation for age-specific cost-effectiveness thresholds that is absent from the static Garber–Phelps framework.

8.4 Status of the Programme

Proposition 1 and Corollary 2 constitute a research programme, not a solved problem. Three conditions must be met before they can be translated into policy instruments.

First, the welfare function $W(H, y)$ must be specified to a degree of precision that allows $W_H(t)$ to be estimated from observable data. The Garber–Phelps framework takes W_H/W_y as the ratio of marginal utilities, estimated via WTP studies; the dynamic framework requires its *path* over the demographic trajectory, which is a substantially harder estimation problem.

Second, the health capital trajectory $H(t)$ in equation (24) is not directly observable. Connecting the theoretical construct to measurable health outcomes (QALYs, disability-adjusted life years, mortality rates) requires a calibrated Grossman model for the relevant population.

Third, the actuarial solvency constraint (13) must be incorporated into the dynamic problem as a path constraint on $B(t)$, not a terminal condition. This requires solving a stochastic control problem over the demographic distribution — a technically demanding extension that lies beyond the scope of this paper but defines the boundary of what would constitute a complete unified framework.

9 Convergence and Synthesis: Mathematical Economics as the Architecture of Health Care Systems

The argument of this paper departs from the simpler convergence narrative. The three traditions do not converge; they are in productive tension. This tension is not a defect of health economics but a symptom of the field's genuine difficulty: it must simultaneously address questions of allocative efficiency, insurance sustainability, and distributive justice, and these questions pull in different directions when formalised with precision.

What the three traditions *do* share is a methodological commitment: that the problems of health economics—pricing risk, allocating scarce resources, sustaining insurance systems across demographic time—cannot be managed through intuition or political judgement alone. Probability theory, decision theory, and actuarial mathematics are not decorative; they are load-

bearing. This conviction is what distinguishes health economics from health policy as a discipline.

Health economics has not merely consumed theoretical developments from economics; it has been a source of innovation. Arrow's 1963 analysis of information asymmetry in medical markets appeared seven years before Akerlof's market-for-lemons paper and anticipated the principal–agent literature by over a decade. Garber's application of Bayesian decision theory to clinical evaluation preceded its widespread adoption in applied statistics. Zweifel's intertemporal insurance models extended optimal contract theory to cover demographic dynamics that the standard contracting literature had not addressed. The traffic of ideas has run in both directions, and health economics has been, in part, a laboratory for economic theory.

The synthesis that emerges is therefore not harmonious but structured. Arrow provides the normative diagnosis: markets fail, institutions are necessary. Grossman provides the micro-foundation: health is produced by individuals over the life cycle, and the demand for medical care is derived from this production. Garber provides the allocation rule: under a budget constraint, rank interventions by ICER relative to the shadow price of the budget. Zweifel provides the financial architecture: insurance systems require actuarial discipline to be sustainable across demographic time. These four frameworks are complementary in identifying what institutions need to do, and incompatible—as Theorem 1 establishes—in specifying how to do all of it simultaneously.

Each of the four frameworks relies on a distinct class of mathematical object: Arrow uses state-preference expected utility and competitive equilibrium theory; Grossman uses dynamic optimisation over a durable capital stock; Garber and Phelps use constrained optimisation and Lagrangian duality; Zweifel uses actuarial present-value calculus and principal–agent theory. Section 8 shows that the cost-effectiveness threshold $\lambda^*(t)$ — the object that links all four — can be derived as an endogenous co-state variable when the frameworks are embedded in a unified dynamic optimisation. The resulting characterisation, equation (23), reveals that the threshold is a forward-looking integral over demographic and fiscal dynamics, not a constant that budget committees can set once and apply indefinitely. The four partial frameworks are not merely in tension; they are four projections of a single, as yet unbuilt, higher-dimensional object.

10 Open Research Frontiers

Five open questions follow directly from the analysis. Each arises at the same boundary: a clean formal framework running up against empirical or normative realities it was not designed to handle.

The most technically immediate concerns the welfare interpretation of moral hazard. Nyman (2001) argues that Pauly's framework misclassifies insurance-induced consumption: when illness strikes, insurance transfers income to the sick, enabling care they could not otherwise

afford. That component is not inefficiency—it is the mechanism through which insurance improves welfare. The income and price effects of insurance are theoretically separable but empirically observationally equivalent; consumption data alone cannot identify which generated an observed increase in utilisation. The RAND experiment (Manning et al., 1987) estimated aggregate demand responses to cost-sharing variation but was not designed to decompose their welfare interpretation. Until that decomposition is achieved, the optimal coinsurance formula rests on an efficiency concept whose normative content is unsettled.

The second problem is the income-sensitivity of cost-effectiveness thresholds. Garber and Phelps (1997) established that the value of a QALY rises with income; NICE, CADTH, and HAS apply uniform thresholds regardless. No paper has produced a theoretically coherent framework for income-differentiated thresholds, because doing so requires specifying a social welfare function with explicit distributional weights—a normative commitment that technocratic agencies are institutionally structured to avoid. Claxton et al. (2015) estimated what the NHS threshold implicitly is from observed budget decisions rather than deriving what it should be. That sidesteps the normative problem rather than resolving it.

Arrow's information asymmetry framework is now structurally incomplete in a different way. The physician–patient asymmetry he identified was durable because grounded in training patients could not replicate. AI diagnostic tools challenge that assumption—but not simply. Information asymmetry in Arrow's model is not about access to information; it is about the capacity to evaluate it. A patient with an AI-generated diagnostic summary still cannot verify whether the algorithm's training data reflects their demographic or how to weight its output against clinical judgement. AI may also create new asymmetries between patients who can interpret algorithmic outputs and those who cannot, stratified along digital literacy rather than medical training. No formal model of these dynamics exists. This gap is not too hard; it requires applying Arrow's analytical structure to a changed information environment, a programme that has not yet been started.

The behavioural challenge to the QALY is foundationally the most severe. Replacing von Neumann–Morgenstern utility with a prospect-theoretic representation does not modify the QALY framework—it collapses it, because the additivity of $Q = \int_0^T u(h(\tau)) e^{-\delta\tau} d\tau$ depends on expected utility axioms. Two responses exist in principle: defend expected utility normatively as the criterion for resource allocation regardless of descriptive violations, or reconstruct health economic evaluation on behavioural foundations from scratch. The first requires a philosophical argument that has not been made explicitly. The second has not been completed. The field occupies an uncomfortable middle position, acknowledging the violations in footnotes while continuing to use the framework operationally.

Zweifel's actuarial solvency model treats both the demographic trajectory and the medical cost trajectory as exogenous—when in reality they interact. Technological progress extends life, increasing longevity pressure on the actuarial balance, but may simultaneously compress morbidity into fewer years near death, reducing cost pressure. Whether ageing populations

are fiscally catastrophic or manageable depends critically on this interaction, which is empirically contested and formally unmodelled. The relevant expertise is distributed across actuarial science, demographic economics, and health technology assessment; no integrated model has assembled it. A complete model would force explicit choices about which cohort bears which costs—analytically inconvenient precisely because they are politically consequential.

What these five problems share is not a common solution but a common structure: the mathematics is not absent but insufficient. Randomised experiments, structural econometrics, and computational simulation are the tools the gaps require. Each is tractable. Each is absent from the theoretical corpus this paper examines.

11 Conclusion

Mathematical health economics is in an unusual position. Its central policy instruments are rigorously derived from theoretical first principles, and the foundations of each have been shown to be empirically fragile. The field knows this. The institutions persist anyway, because no alternative has achieved comparable tractability at scale.

Theorem 1 reframes the situation. What the literature has treated as a collection of separately noted tensions turns out to be a single structural result: the three incompatibilities flow from the same set of regularity conditions that any empirically plausible health system must satisfy. But the incompatibility result is the beginning of the analysis, not its conclusion. Section 8 develops what comes next.

The dynamic cost-effectiveness threshold $\lambda^*(t)$, defined by equation (23), is the key missing object. It endogenises the central policy instrument of health technology assessment — the cost-effectiveness threshold — by expressing it as the present value of future marginal health benefits discounted at the user cost of health capital. This connects Garber's allocation criterion, Grossman's investment theory, and Zweifel's demographic finance into a single forward-looking integral. Corollary 2 derives its first substantive implication: uniform thresholds are dynamically inconsistent with ageing populations, and the direction of the bias is theoretically determinate.

What remains is threefold: estimating the path of $W_H(t)$ from observable data; calibrating the health capital trajectory $H(t)$ from measurable health outcomes; and incorporating actuarial solvency as a path constraint in the stochastic control problem. These are tractable research tasks, not open-ended conjectures. The dynamic framework converts the incompatibility of Theorem 1 from an obstacle into a map.

References

- Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53(5), 941–973.
- Grossman, M., 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80(2), 223–255.
- Garber, A.M., Phelps, C.E., 1997. Economic foundations of cost-effectiveness analysis. *Journal of Health Economics* 16(1), 1–31.
- Zweifel, P., Breyer, F., Kifmann, M., 2009. *Health Economics*, 2nd ed. Springer, Berlin.
- Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P.C., Sculpher, M., 2015. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technology Assessment* 19(14), 1–503.
- Arrow, K.J., 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Économétrie* 11, 41–47.
- Akerlof, G.A., 1970. The market for ‘lemons’: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3), 488–500.
- Rothschild, M., Stiglitz, J.E., 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90(4), 629–649.
- Garber, A.M., 2000. Advances in cost-effectiveness analysis of health interventions. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, Vol. 1A. Elsevier, Amsterdam, pp. 181–221.
- Garber, A.M., 2004. Cost-effectiveness and evidence evaluation as criteria for coverage policy. *Health Affairs* W4, 284–296.
- von Neumann, J., Morgenstern, O., 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.
- Pliskin, J.S., Shepard, D.S., Weinstein, M.C., 1980. Utility functions for life years and health status. *Operations Research* 28(1), 206–224.
- Drummond, M.F., Sculpher, M.J., Claxton, K., Stoddart, G.L., Torrance, G.W., 2015. *Methods for the Economic Evaluation of Health Care Programmes*, 4th ed. Oxford University Press, Oxford.
- Torrance, G.W., 1986. Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics* 5(1), 1–30.

- Johannesson, M., 1996. *Theory and Methods of Economic Evaluation of Health Care*. Kluwer Academic Publishers, Dordrecht.
- Viscusi, W.K., Aldy, J.E., 2003. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27(1), 5–76.
- Pauly, M.V., 1968. The economics of moral hazard: Comment. *American Economic Review* 58(3), 531–537.
- Nyman, J.A., 2001. *The Theory of Demand for Health Insurance*. Stanford University Press, Stanford.
- Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., Leibowitz, A., Marquis, M.S., 1987. Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review* 77(3), 251–277.
- Zweifel, P., Manning, W.G., 2000. Moral hazard and consumer incentives in health care. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, Vol. 1A. Elsevier, Amsterdam, pp. 409–459.
- Papanicolas, I., Woskie, L.R., Jha, A.K., 2018. Health care spending in the United States and other high-income countries. *JAMA* 319(10), 1024–1039.
- Weinstein, M.C., Stason, W.B., 1977. Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine* 296(13), 716–721.
- Bleichrodt, H., Quiggin, J., 1997. Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty* 15(2), 151–165.
- Murphy, K.M., Topel, R.H., 2006. The value of health and longevity. *Journal of Political Economy* 114(5), 871–904.
- Zeckhauser, R., 1970. Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2(1), 10–26.
- Ehrlich, I., Chuma, H., 1990. A model of the demand for longevity and the value of life extension. *Journal of Political Economy* 98(4), 761–782.



UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
44/50 DŁUGA ST.
00-241 WARSAW
WWW.WNE.UW.EDU.PL
ISSN 2957-0506