Moral Hazard under Liquidity Constraints*

Keith Marzilli Ericson†, Johannes G. Jaspersen‡, Justin Sydnor§

February 10, 2026

Abstract

Insurance-induced spending — moral hazard—is often assumed to be inefficient. Using a graphical framework and formal results, we incorporate liquidity constraints into standard models and show that the resulting financing benefit of insurance makes part of moral hazard efficient. Simulations show economically substantial net benefits of moral hazard. Revisiting the Oregon Medicaid Experiment, we estimate liquidity-constrained individuals value Medicaid over twice as much as unconstrained individuals, and over $\$1,000$ above Medicaid's net costs. Finally, simulations show how liquidity constraints increase the welfare benefits of higher insurance coverage and the value of vertical choice in insurance.

*We acknowledge funding from the Boston University Questrom School of Business. None of the authors has a conflict of interest with this work. We thank Angie Acquatella, Zarek Brot, Stuart Craig, Cecilia Diaz-Campo, Amy Finkelstein, Daniel Gottlieb, Nathaniel Hendren, Reiner Leidl, Victoria Marone, Chad D. Meyerhoefer, Wanda Mimra, Philip Mulder, Joseph Newhouse, John Nyman, Casey Rothschild, Dan Sacks, and Andre Veiga as well as participants at the 2024 annual meeting of the American Society of Health Economists, the 2025 annual meeting of the Risk Theory Society, the University of Wisconsin, Harvard University, Lehigh University, Northeastern University, Toulouse School of Economics, University of Rochester, the Becker-Friedman Health Economics seminar, and the CEAR/ MRIC Behavioral Insurance Workshop for helpful comments.

†Boston University Questrom School of Business and NBER, email: kericson@bu.edu

‡LMU Munich School of Management, LMU Munich, email: jaspersen@lmu.de

§Wisconsin School of Business, University of Wisconsin – Madison, email: justin.sydnor@wisc.edu

1 Introduction

The change in spending resulting from health insurance coverage is typically referred to as moral hazard.[1] This empirical definition—a change in spending—is often paired with a welfare interpretation: moral hazard is considered "excess" spending that individuals value less than its full cost due to a distortion in incentives (Pauly, 1968). The dominant modeling frameworks used in the analysis of the value of health insurance imply that moral hazard lowers welfare (Einav et al., 2013; Finkelstein et al., 2019; Marone and Sabety, 2022).

We revisit this interpretation of moral hazard, building on insights from De Meza (1983), Nyman (1999a), and Chetty (2008) that moral hazard can actually be efficient and welfare increasing. We examine liquidity constrained individuals, for whom insurance provides both a risk-protection benefit and access to financing: while insurable expenses are often concentrated in a particular consumption period, the cost of insurance via premiums is typically spread across multiple consumption periods. Insurance can thus smooth expenditures across time, not merely across states (e.g., Ericson and Sydnor, forthcoming; Gross et al., 2022). This financing property can quantitatively and qualitatively change patterns of insurance demand and spending.

Accounting for liquidity constraints is important, particularly for evaluating expansions of insurance to low-income individuals. It is well-established that many individuals face liquidity constraints (either the inability to borrow or high interest rates) and have high marginal-propensities to consume out of income shocks (Parker et al., 2013; Colarieti et al., 2024). As a result, macroeconomic models often treat a sizable share of individuals as living essentially hand-to-mouth, consuming all liquid resources in a given period (Kaplan et al., 2014; Lee and Maxted, 2023; Aguiar et al., 2025). Moreover, because of holdings in illiquid assets, hand-to-mouth behavior can describe even wealthy, optimizing individuals (Kaplan et al., 2014). Liquidity constraints have been shown to affect healthcare utilization (Gross et al., 2022), and healthcare providers in the U.S. increasingly demand cost-sharing payments before delivering care (e.g., Washington, 2025). Further, even if people ultimately have access to liquidity, if they have behavioral biases that lead them to act "as if" they were hand-to-mouth (Olafsson and Pagel, 2018; Lee and Maxted, 2023), those perceived liquidity constraints will distort spending when uninsured.[2]

We analyze the welfare implications of ex-post moral hazard in light of liquidity constraints. Our setting is health insurance, which has been the focus of a substantial body of literature on moral hazard (Einav and Finkelstein, 2018), but the model is broadly applicable to various types of insurance. Standard models of ex-post moral hazard in (health) insurance are one-period models in which premiums, cost-sharing, and consumption all occur in the same consumption period. Our

[1] For instance, Einav and Finkelstein (2018) write "we follow decades of health insurance literature and use the term "moral hazard" to refer to the responsiveness of healthcare spending to insurance coverage". For a discussion of the history of the term and alternative definitions, see Rowell and Connelly (2012).

[2] In our model, we assume individuals are optimizing. However, new issues may arise if financial stress itself affects decision-making, as in Sergeyev et al. (2025).

innovation, which closely follows the approach in Ericson and Sydnor (forthcoming), is to introduce multiple underlying consumption periods within the insurance period. The cost of insurance via premiums is spread equally across all N consumption periods within the insurance policy term. Uninsured medical expenses, however, are spread across $K \leq N$ periods. In this way we introduce a portable extension of the standard model, in which $K = N$ embeds the standard model that implicitly assumes perfect liquidity within the insurance-contract duration, while $K = 1$ is the full hand-to-mouth case, in which all uninsured expenses are borne in a single consumption period. Ericson and Sydnor (forthcoming) use this framework to examine the value of insurance under liquidity constraints in the absence of moral hazard. The current paper explicitly focuses on moral hazard and the endogenous decision of medical spending.

We also innovate on existing discussions of moral hazard by explicitly defining the efficient level of medical spending for a given level of insurance coverage, building on insights from De Meza (1983) and Nyman (1999a).[3] We define efficient spending for a given level of insurance as the medical spending the individual would choose if they maximized ex-ante utility, fully internalizing how medical spending would affect insurance premiums. This represents the level of spending they would contractually commit to, recognizing they have to pay for the expected amount via premiums, if it were possible to create a binding contract.[4] This efficient benchmark can then be compared to two alternative scenarios: the level of medical spending chosen when uninsured, and the level chosen when insured but taking the premiums for insurance as given. The chosen level of spending under insurance will be higher than the efficient level, because premiums are fixed for a contract period and insureds do not internalize the impact additional spending has on premiums. The focus of this paper is that the uninsured level of spending will generally be below the efficient level and especially so for those with liquidity constraints.

Our framework allows us to present simple definitions decomposing moral hazard into both inefficient and *efficient* components. We develop intuitive graphical depictions of the model that allow us to illustrate the welfare impact of moral hazard and how it depends on liquidity constraints. This framework makes it clear how, with familiar structural assumptions about utility of consumption and health, the welfare impacts of efficient and inefficient moral hazard can be quantified using the same data on medical spending chosen under different levels of insurance that have previously been used to evaluate the impact of moral hazard in empirical studies (e.g., Finkelstein et al., 2019).

---

[3] Both De Meza (1983) and Nyman (1999a) discuss the potential efficiency of moral hazard from health insurance. De Meza focuses on an insurance contract that pays a fixed income when sick, while Nyman considers traditional health insurance that covers a portion of the costs and hence lowers the price of care. Our approach is similar to Nyman, though as we discuss in Section 2, our way of defining the efficient level of moral hazard is different.

[4] The efficient level of spending defined this way at full insurance will be the *first-best* level of spending. Defining the more general notion of efficient spending conditional on the insurance level allows us to explore moral hazard across a range of insurance coverage. Acquatella and Marone (2025) leverage the first-best level of spending in their analysis of the value of moral hazard without liquidity constraints in a model with state-dependent preferences.

Our theory first considers the standard intuition that moral hazard is inefficient in the absence of liquidity constraints. For this analysis we focus on the case when the loss is certain to happen (or has already happened). With perfect liquidity, optimization implies that when uninsured, individuals are indifferent between a dollar more of medical spending and a dollar more of consumption on the margin. Adding insurance lowers the price individuals face, causing them to increase expenditures, but this expenditure is not valued by individuals with perfect liquidity at its full cost. With perfect liquidity the efficient level of spending is the uninsured amount regardless of the level of insurance. As a consequence, the standard result holds for those with perfect liquidity and certain losses: increasing insurance lowers welfare because the additional spending is inefficient.

For liquidity-constrained individuals, the results are quite different. The standard intuition no longer holds. With liquidity constraints, optimization implies that, when uninsured, individuals are indifferent on the margin between a dollar more of out-of-pocket medical spending and a dollar of consumption *in that period*. However, they strictly prefer forgoing a dollar spread across multiple consumption periods compared to forgoing a dollar in a single consumption period of high medical care consumption. Thus, an additional dollar of medical spending in the sick period will be more valuable than the cost of financing that spending with a dollar of premiums (or equivalently in taxes) spread across periods. As a result, for liquidity-constrained individuals, the efficient level of spending increases with the level of insurance. The individual will choose to spend more than the efficient level, since insurance distorts prices. Yet, the benefits of increases in efficient moral hazard can outweigh the inefficient moral hazard. The net effect can be that an increase in spending due to an expansion of insurance, i.e., the moral hazard spending, increases rather than reduces welfare.

We show in numerical simulations that the net welfare impacts of moral hazard can be significantly positive for individuals living hand-to-mouth. Since moral hazard can be net-positive for those with liquidity constraints, the optimal level of insurance can also be much higher for them.[5] In standard theory, when demand is more responsive to price – when there is more moral hazard – the level of insurance should be lower (Zeckhauser, 1970; Pauly and Blavin, 2008). However, if environments with more moral hazard are settings where people have more liquidity constraints, the standard intuition could be backward. Empirically, the degree of moral hazard is often identified through the insurance-coverage elasticity of demand for medical care. We show formally that the presence of liquidity constraints changes how this parameter should be interpreted and that with more liquidity constraints, a given observed elasticity implies a higher optimal insurance coverage.

We also characterize the efficient level of spending and show how it relates to observed moral hazard for individuals with and without liquidity constraints. We show that the share of observed moral hazard that is efficient is substantially higher for liquidity-constrained individuals, and for them the efficient level of spending can be much closer to the insured level than the uninsured level. This matters because insurance contracts and policy interventions often rely on supply-

[5] This is also reflected in our formal propositions. Liquidity-constrained individuals have a larger range of insurance coverage levels with a positive net value of moral hazard. This lets us show that the welfare-maximizing level of insurance coverage is higher in the presence of liquidity constraints than in the canonical case of perfect liquidity.

side utilization controls to limit spending, not just demand-side cost-sharing (Ellis and McGuire, 1986, 1990; Dranove et al., 2021; Layton et al., 2022). Determining the optimal level of such restrictions requires a view on how much observed moral hazard reflects inefficient overuse versus efficient increases in care that arise because insurance relaxes liquidity constraints. If policymakers interpret most observed moral hazard as waste, they will view uninsured spending as close to the efficient benchmark. This may lead them to design supply-side restrictions to mimic uninsured utilization—potentially restricting access to treatments that are socially valuable but too costly for liquidity-constrained individuals to absorb.

Our framework also makes it possible to revisit an important insight from Nyman (1999a): even in the absence of liquidity constraints, moral hazard induced by health insurance is not entirely inefficient. Nyman highlights that there is an income effect from insurance when the probability of illness is less than one. The obvious benefit of insurance comes from the ex-ante value of reducing spending risk into a sure premium. However, ex-post, when an individual is sick, they are also receiving a transfer from those who did not get sick. The optimal level of spending will be higher than when uninsured due to the income value of that transfer.[6] Our definitions of efficient and inefficient moral hazard and our graphical framework make it possible to intuitively visualize this "Nyman income effect" and to quantify the welfare value of moral hazard given this income effect.

We compare how liquidity constraints and the Nyman income effect impact the value of moral hazard in simulations. The Nyman income effect depends more strongly on the probability of loss than the liquidity effect of insurance does. In our simulations, for those without liquidity constraints the Nyman income effect is generally not strong enough to make the net value of moral hazard positive.[7] Those with liquidity constraints, however, get both a liquidity benefit and the Nyman income effect from insurance. For them, the net value of moral hazard is significantly positive in many scenarios.

We use our model in the empirical setting of Finkelstein et al. (2019) to estimate of the value of Medicaid, the public health insurance program for low-income Americans. We use a simple binary-loss framework, focusing on how incorporating liquidity constraints affects the interpretation of moral hazard and the welfare consequences of Medicaid expansion. Consistent with the original Finkelstein et al. (2019) findings, we estimate that individuals with perfect liquidity who can fully spread any uninsured medical spending throughout the year would value the moral hazard spending induced by Medicaid at only half its uninsured cost and overall would value Medicaid at less than the net cost of providing Medicaid to them. However, if Medicaid recipients are hand-to-mouth

[6] This "Nyman income effect" arises via price changes induced by insurance and is different than the effect of an increase in permanent income on uninsured spending decisions. De Meza (1983) discussed efficient moral hazard but in the context of response to an increase in permanent income.

[7] This finding is naturally sensitive to the parameters of our simulations. Our simulations focus on modeling the overall value of health insurance over a range of parameters broadly similar to those in studies such as Finkelstein et al. (2019). It is quite plausible that for some types of coverage, especially protection against rare diseases for which there is little value of (or possibility of) treatment at lower levels of spending, the net value of moral hazard can be positive, given the Nyman income effect.

types and must absorb uninsured medical costs over only two months, we get dramatically different results. Under these liquidity constraints, Medicaid recipients would value Medicaid by $\$1,000$ more than the net cost of providing Medicaid to them. We estimate that the largest component of Medicaid's value to the liquidity constrained comes from their valuation for the moral hazard spending. They value it at 150% of its expected cost, implying a significant net positive welfare impact of moral hazard.[8]

In a final illustration of the applications of our model, we consider the implications of this framework for situations where plan selection interacts with moral hazard (e.g., Einav et al., 2013; Azevedo and Gottlieb, 2017; Marone and Sabety, 2022). We adapt parameterized models from prior literature and conduct simulations to investigate how conclusions change if individuals are liquidity constrained. We show first that if there is underlying heterogeneity in liquidity constraints but the econometrician does not model liquidity, it will bias estimates of the moral-hazard-elasticity parameter in standard models. This can dramatically change the estimated optimal level of coverage for the population: large estimated moral-hazard elasticities call for lowering insurance coverage, while moral-hazard responses stemming from liquidity constraints call for more insurance coverage. We show further that while vertical choice in insurance may be inefficient in standard frameworks (Marone and Sabety, 2022), there can be substantial welfare value to vertical choice if heterogeneity in moral hazard arises primarily from variation in liquidity.

Our paper contributes to a small literature that has focused on the efficient components of moral hazard responses to insurance (De Meza, 1983; Nyman, 1999a,b, 2003; Chetty, 2008; Acquatella and Marone, 2025). The central findings of our paper are most closely related to Chetty (2008), who examines how unemployment durations respond to unemployment insurance benefits. Chetty's key insight is that while a portion of these responses are an inefficient reaction to distorted incentives, part of the response can be efficient if it results from overcoming liquidity constraints. Our paper echoes this point and provides a new way of visualizing and developing intuition for it. The primary difference between our paper and Chetty (2008) lies in how we model and quantify the liquidity value of insurance. Chetty derives a formula for the optimal benefit levels that depends on two different elasticities: a response to lump-sum cash and a response to incentives on the margin. While this approach could be conceptually extended to analyze moral hazard in health insurance, Chetty and Finkelstein (2013) highlight that it requires developing quasi-experimental strategies for identifying lump-sum payments in health insurance that identify the liquidity effect. Perhaps in part due to this challenge, the liquidity effect has not been incorporated into the analysis of moral hazard in health insurance. Our approach is an extension of the models currently used in

[8] A related working paper (Mukherjee et al., 2024) finds that Medicaid expansion did not substantially reduce the variability of measured consumption for the likely eligible population, implying that Medicaid has low insurance value. Under liquidity constraints, the smoothness of consumption across time matters in addition to the overall variability of consumption, but even if providing Medicaid has little impact on consumption patterns, our results show that Medicaid's benefits via moral hazard can be large. If uninsured individuals limit their medical spending because it is costly to reduce consumption, that can limit the apparent consumption-smoothing impacts of insurance but implies the medical spending financed via insurance is highly valuable.

the health insurance literature (e.g., Finkelstein et al., 2019), which means that the same types of data and assumptions used in recent empirical analyses of moral hazard can be extended to incorporate liquidity constraints using our framework. A secondary point of differentiation between our work and Chetty (2008) is that Chetty's framework starts from the presumption that the individual is unemployed and abstracts from the risk-protection benefits of insurance, while we do not. Incorporating the uncertainty in whether the sick state will happen allows us to identify and distinguish the Nyman income effect from the liquidity effect of health insurance and to examine the magnitude of the net moral hazard value relative to the risk-protection and financing value of insurance.

2 Model

2.1 Layout

In our model, individuals live N periods and have lifetime utility consisting of consumption and health components, $U = \sum_{t=1}^{N} \delta^{t-1} \left(u(c_t) + h(m_t, \theta_t)\right)$.[9] We assume the exponential discount factor $\delta = 1$ throughout to highlight the role of liquidity constraints apart from time preference; with horizons less than a year, δ should be near 1 (Ericson and Laibson, 2019).

Consumption utility in each period t is given by $u(c_t)$, with $u(\cdot)$ continuous, $u' > 0$ and $u'' < 0$. Individuals receive total income y, paid in equal installments each period. Utility from health is separable from consumption and is given each period by $h(m_t, \theta_t)$, where θ_t indexes the individual's health state and m_t indexes the level of medical spending in period t.

We make a set of simplifying assumptions for our baseline model. We assume there are two health states: $\theta = sick$ and $\theta = healthy$. With probability $1 - \pi$ the individual is healthy for all N periods. With probability π the individual is sick in one of the N periods. We assume that the realization of the health state occurs at the beginning of the first period, so the individual knows whether and when they will be sick before choosing consumption.[10]

In the healthy state, health is unaffected by medical spending and $h(m, healthy)$ is normalized to be zero. In the sick state, medical spending initially generates a positive but decreasing marginal benefit. We suppress the θ argument when sick and write that $h'(m) > 0$ and $h''(m) < 0$ for all values of m smaller than $\bar{m}$. The function gives benefits of medical spending up to a point. Once $\bar{m}$ is reached, more medical spending has a negative effect on health such that $h'(m) < 0$ and

[9] This model and exposition build on Ericson and Sydnor (forthcoming), adding the possibility of choosing medical spending as in Finkelstein et al. (2019). In contrast to the Finkelstein et al. (2019) model, we assume N consumption periods in a policy period rather than just one.

[10] This simplifying assumption eliminates the need to consider more complicated dynamics related to the potential timing of when a sick state might arise during the N periods that would affect consumption decisions for individuals that can borrow and save across consumption periods. For individuals that are fully constrained to live hand-to-mouth and can neither borrow nor save, this assumption can be relaxed without complicating the model. Ericson and Sydnor (forthcoming) show that this assumption – termed perfect foresight – is necessary for an individual with no liquidity constraints to be represented with a standard static expected utility model.

$h''(m) < 0$. This is in line with the models of Einav et al. (2013) and Marone and Sabety (2022) and captures the idea that after reaching a certain health state additional medical spending is either useless (but still leads to detriments due the associated hassle costs) or actively harmful.

As a result of these assumptions, an individual's utility is simply $U = \sum_{t=1}^{N} u(c_t)$ if healthy and $U = h(m) + \sum_{t=1}^{N} u(c_t)$ if sick. The individual chooses their consumption vector c and the level of medical spending m after observing the realization of the health state subject to budget constraints. This choice process is described in more detail below as we describe two types of individuals varying in their liquidity constraints.

Before period 1, the individual enters into an insurance plan characterized by a total insurance premium p and proportional insurance level α. That is fraction α of medical spending is covered by insurance, leaving the individual with out-of-pocket costs $(1-\alpha)m$. Note that the level of insurance α will affect the level of medical spending optimally chosen by the individual – this allows for the possibility of moral hazard.

We assume that premium payments are made smoothly over time, such that they pay a $1/N^{\text{th}}$ of the premium in each period.[11] We assume that premiums are actuarially fair such that $p = \pi\alpha m$ and reflect the endogenously chosen level of medical spending in the sick state.[12] As in most models incorporating the possibility of moral hazard, the individual will choose medical spending taking the level of the premium as fixed and not internalizing the impact of medical spending on premiums.

2.2 Types of Individuals

Individuals with "perfect liquidity" can allocate their total income toward consumption, premium payments and medical spending across periods, and simply must satisfy a lifetime budget constraint; they can costlessly borrow from future income, and their savings earn no rate of return.

In contrast, individuals with liquidity constraints live hand-to-mouth – they can only move resources between periods to a limited degree: they have to pay any out-of-pocket medical spending spread over $K_i < N$ consumption periods.[13]

We consider individuals who differ in their liquidity constraints, indexed by $K_i \in \{1, 2, \ldots, N\}$, the number of consumption periods over which out-of-pocket medical spending can be spread. For an individual who can spread the spending across K_i periods, the ex-ante expected utility is given

[11] This assumption of smooth premium payments matches the empirical situation that is typical for health insurance contracts in the United States. It could also represent public insurance financed via income taxes. See Ericson and Sydnor (forthcoming) for a discussion of alternative premium-payment timing.

[12] This assumption that premiums reflect the chosen level of medical spending aligns with the model of Nyman (2003) and could imply either that they are set with correct expectations of how individuals will respond to their insured incentives or could be thought of as a longer-run equilibrium result where premiums and spending decisions adjust to be mutually consistent.

[13] The stark hand-to-mouth case facilitates developing theoretical results. As shown in Ericson and Sydnor (forthcoming), the same insights translate to a more computationally difficult model in which individuals can borrow at high interest rates.

by:

$$EU_i = \pi \left[(N - K_i) u \left(\frac{y-p}{N} \right) + K_i u \left(\frac{y-p}{N} - \frac{(1-\alpha)m(\alpha)}{K_i} \right) + h(m(\alpha)) \right] + (1-\pi) N u \left(\frac{y-p}{N} \right). \quad (1)$$

Results differ for types of individuals. The first type, denoted HTM for hand-to-mouth and has $K_{HTM} < N$. $K_i = 1$ describes the extreme case in which the individual receives a medical treatment in a given consumption period and has to pay for it immediately. $K_i > 1$ but less than N can appear if the medical treatment and associated bills are spread over several periods, or the providers allow for a payment plan of the bills.

The second type is denoted PL for perfect liquidity. Setting $K_{PL} = N$ in Equation (1) yields the following utility function for an individual with perfect liquidity, as their optimal decision will be to perfectly smooth consumption across all periods. Thus, ex-ante expected utility is:

$$EU_{PL} = \pi \left[N u \left(\frac{y - (1-\alpha)m(\alpha) - p}{N} \right) + h(m(\alpha)) \right] + (1-\pi) N u \left(\frac{y-p}{N} \right). \quad (2)$$

This preference functional is conceptually equivalent to static expected utility models used in prior work on health insurance and moral hazard, including Finkelstein et al. (2019).

Hereafter, we drop the i subscript when the notation is clear without it.

2.3 Defining Moral Hazard

We distinguish between the *chosen* and *efficient* level of medical spending for a given amount of insurance coverage. The level of medical spending chosen when sick to maximize ex-post utility for an individual of type i with insurance coverage α is $m_i^*(\alpha)$. The individual will select this amount of spending taking the premium cost of insurance as given. When an individual has no insurance, their chosen medical spending is denoted $m_i^*(0)$.

The efficient (i.e., optimal) level of medical spending is denoted by $m_i^E(\alpha)$. This is the level of medical spending that individuals of type i, taking the level of insurance α as given, would commit to *ex-ante*. For the efficient level, the individuals would internalize the full effect of medical spending on their premiums. The efficient level of medical spending, for a given level of insurance, is thus the level of m that maximizes expected utility subject to the constraint that $p(m, \alpha) = \pi \alpha m$. That is, $m_i^E(\alpha) = \arg\max_m EU_i(m, p(m, \alpha), \alpha)$. Note that the level of efficient spending will depend on the level of insurance as the utility costs of spending will depend on how it is financed.[14] In the next section we explore the first order conditions that govern the efficient and chosen levels of spending.

Our definition of efficient spending differs from the definition used by Nyman (1999a) and summarized in Nyman et al. (2018). Nyman defines efficient spending as the level that would

[14] This distinguishes the efficient level of spending from a *first-best* solution to the decision problem. In our setting without transaction cost, the first-best contract would imply full risk transfer and an *ex-ante* commitment to a given level of spending when sick. In terms of the model, the first-best solution appears for $\alpha = 1$ and $m^E(1)$ which is equivalent for PL and HTM individuals, because $\alpha = 1$ removes all liquidity constraints. We use an efficient level of spending conditional on the level of insurance because it allows us to distinguish between efficient and inefficient moral hazard for any insurance contract.

arise from receiving a lump-sum transfer when sick equal to the amount the individual would have spent under insurance (see also Chetty, 2008). In Appendix B we present a deeper discussion and derivation of the differences between our measure and the lump-sum measure. Setting the lump sum amount equal to the amount the individual would spend under insurance will not achieve the efficient level of spending as we define it. The problem is that the total amount being spent when insured is inefficient and hence the lump-sum transfer is too large. Lump-sum contingent transfers could be used in theory to achieve the efficient level of spending, but the amount of the lump sum transfer would need to be chosen to yield $m_i^E(\alpha)$ as the result of the individual's maximization problem. We believe our definition of efficient spending is the correct one, since it is derived from the full internalization of the utility cost of spending given the level of insurance.

Having set up these definitions, we can now define:

1. Observed moral hazard: $MH_i(\alpha) = m_i^*(\alpha) - m_i^*(0)$
2. Efficient moral hazard: $MH_i^E(\alpha) = m_i^E(\alpha) - m_i^*(0)$
3. Inefficient moral hazard: $MH_i^I(\alpha) = m_i^*(\alpha) - m_i^E(\alpha)$

The first empirical object $MH_i(\alpha)$ is the total observed change in medical spending when an individual has insurance level α versus when they are uninsured. The observed moral hazard can be disentangled into an efficient ($MH_i^E(\alpha)$) and inefficient ($MH_i^I(\alpha)$) component. Note that $MH_i(\alpha) = MH_i^E(\alpha) + MH_i^I(\alpha)$. The efficient component of moral hazard is the change in spending that the individual would optimally choose, while the inefficient component of moral hazard is the "excess" spending.

2.4 Defining the Value of Insurance

These concepts allow us to decompose the value of insurance into a component resulting from moral hazard, and a separate component from its effect on risk and financial smoothing. The total value of insurance for an individual of type $i \in \{PL, HTM\}$ at coverage level α is the change in expected utility relative to no insurance:

$$V_i(\alpha) = EU_i(\alpha, m_i^*(\alpha)) - EU_i(0, m_i^*(0)), \tag{3}$$

where each of the expected utility terms are evaluated at the actuarially fair premiums consistent with the chosen level of medical spending ($m_i^*(\alpha)$ and $m_i^*(0)$, respectively).

This total value can be split into the moral hazard value and a "risk and financing value" such that $V_i(\alpha) = V_i^{MH}(\alpha) + V_i^{RF}(\alpha)$, where

$$V_i^{RF}(\alpha) = EU_i(\alpha, m_i^*(0)) - EU_i(0, m_i^*(0)) \text{ and} \tag{4}$$

$$V_i^{MH}(\alpha) = EU_i(\alpha, m_i^*(\alpha)) - EU_i(\alpha, m_i^*(0)). \tag{5}$$

The risk and financing value V_i^{RF} is the change in utility from buying actuarially fair insurance, but holding fixed the medical spending at the uninsured level $m_i^*(0)$. This entails using insurance to reduce risk and finance spending but not to change spending.[15]

The moral hazard value $V_i^{MH}(\alpha)$ is the change in utility from the individual choosing medical spending ex-post with insurance ($m_i^*(\alpha)$) versus having medical spending fixed at the uninsured level $m_i^*(0)$. This value can be further decomposed into the value of efficient moral hazard, V_i^{MHE}, and the value of inefficient moral hazard, V_i^{MHI}, as follows:

$$V_i^{MHE}(\alpha) = EU_i(\alpha, m_i^E(\alpha)) - EU_i(\alpha, m_i^*(0)) \text{ and} \tag{6}$$

$$V_i^{MHI}(\alpha) = EU_i(\alpha, m_i^*(\alpha)) - EU_i(\alpha, m_i^E(\alpha)). \tag{7}$$

Again, $V_i^{MH}(\alpha) = V_i^{MHE}(\alpha) + V_i^{MHI}(\alpha)$. From the definitions, it is clear that $V_i^{MHE} \geq 0$ and $V_i^{MHI} \leq 0$. While it is typically assumed that $V_i^{MH}(\alpha)$ is negative (and thus $-V_i^{MHI} > V_i^{MHE}$), our next section will show that it can be positive. In particular, we make clear that it is more likely to be positive when individuals are liquidity constrained and argue that this is the reason why the value of insurance will be positive for a larger range of possible coinsurance levels.

3 Model Results

We begin with an intuitive discussion of how the analysis of moral hazard changes when we consider the liquidity and income-effects of insurance. We develop a set of simple graphical depictions that make it possible to visualize both the moral-hazard and welfare implications of insurance. We then present results of numerical simulations that help to highlight the potential magnitude of the differences in insurance value and specifically the impact of moral hazard. In the third subsection, we present formal theoretical propositions under more general conditions with proofs provided in the appendix. Some readers with less interest in the technical formalities may choose to skip that subsection without losing intuitions for the core results. The final subsection connects our finding to an often observed empirical object: the coverage elasticity of demand for medical care.

3.1 Graphical Analysis and Intuition

We begin with an intuitive discussion of the key results. We highlight the following key points:

1. The liquidity benefits of insurance lead part of observed moral hazard to be efficient for hand-to-mouth agents. There can be positive moral hazard value from insurance for hand-to-mouth agents even when the sick state is sure to occur.

[15] As Ericson and Sydnor (forthcoming) discuss and we highlight in the next section, for those with perfect liquidity, insurance has only a risk-reducing benefit. It has an additional financing benefit for those with liquidity constraints.

2. When the sick state is not sure to occur, there is also an income effect of insurance ("Nyman income effect"). This leads part of observed moral hazard to be efficient for both perfectly liquid and hand-to-mouth agents.

3.1.1 Illustrating the Liquidity Effect Using the Sure-loss Case ($\pi = 1$)

When losses are sure to occur ($\pi = 1$), the perfect-liquidity case conforms to the standard economic intuition introduced by Pauly (1968). For those with perfect liquidity, the efficient level of spending when $\pi = 1$ is the level that would be chosen when uninsured. To see this, we recall the expected utility for the perfect liquidity case from Equation (2) and start by considering the efficient level of spending. Since premiums are defined by $p = \pi\alpha m$ and $\pi = 1$ here, the equation reduces to:

$$EU_{PL} = Nu\left(\frac{y-m}{N}\right) + h(m). \tag{8}$$

This equation does not depend on the level of insurance α, since any increase in spending must be paid for fully either via uninsured spending or insurance premiums. As such, the first order condition implicitly determining the optimal level of spending when uninsured ($m^*_{PL}(0)$) and the efficient level of spending ($m^E_{PL}(\alpha)$) is the same and given by:

$$FOC^E_{PL} = FOC^0_{PL} : u'\left(\frac{y-m}{N}\right) = h'(m). \tag{9}$$

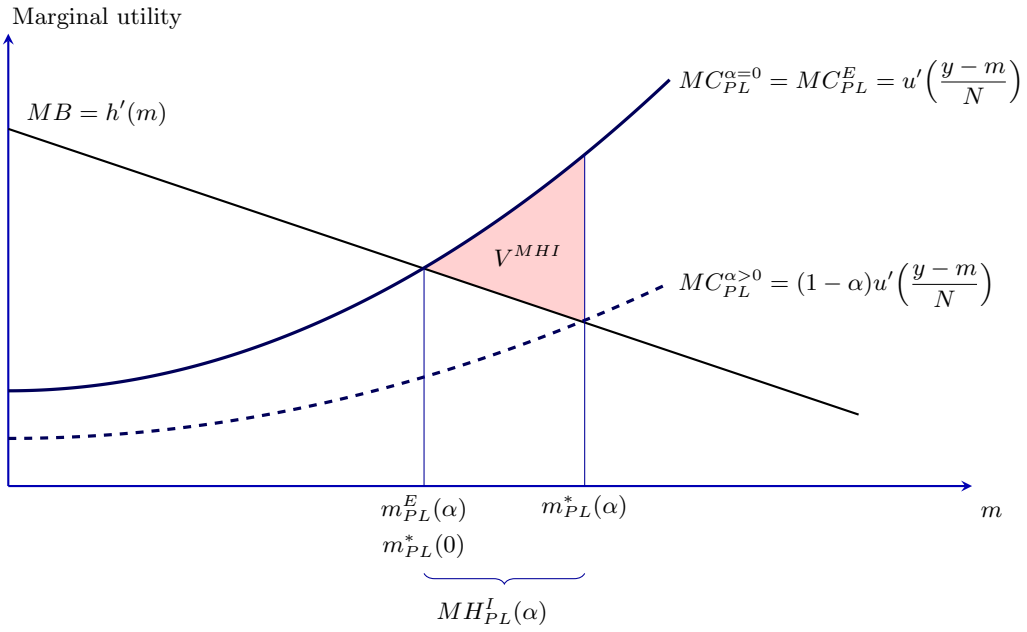
Insurance leads to inefficient moral hazard, as the individual takes the premium as fixed when choosing their desired level of spending ($m^*_{PL}(\alpha)$), which is defined by the following first order condition:

$$FOC^*_{PL} : (1-\alpha)u'\left(\frac{y-m}{N}\right) = h'(m). \tag{10}$$

Figure 1 gives a graphic illustration of this standard case of inefficient moral hazard for an individual with perfect liquidity. The vertical axis is the marginal utility cost and benefit of additional medical spending. The horizontal axis is the level of medical spending. The marginal benefit (MB) of spending is given by $h'(m)$, which is downward sloping given our assumptions on $h(m)$.

We then plot the marginal costs (MC) of medical spending, which is the marginal consumption utility given the level of medical spending (i.e., the left-hand side of the first-order conditions in Equations (9) and (10)). These MC lines slope upward due to the concavity of the utility function ($u'' < 0$). The line $MC^{\alpha=0}$ is the marginal utility cost of foregone consumption with medical spending m when uninsured. Next, MC^E is the marginal cost of additional medical spending accounting for the additional premiums induced by that spending. Here it lies directly on top of $MC^{\alpha=0}$, as noted above. The uninsured and efficient levels of spending coincide: $m^*(0) = m^E(\alpha)$. The dashed line, $MC^\alpha > 0$, is the marginal utility cost of foregone consumption through out-of-pocket costs given insurance α at medical spending level m. It is shifted down, since insurance

Figure 1 – Standard Analysis of Moral Hazard (Perfect Liquidity)



Note: Schematic illustration. Assumes perfect liquidity $K = N$ and sick state happens for certain $\pi = 1$. m is medical spending in the sick state of the world.

means out-of-pocket costs are lower for any given amount of medical spending. As a result, the medical spending $m_{PL}^{*}(\alpha)$ chosen when insured is higher, with the additional spending representing inefficient moral hazard ($MH_{PL}^{I}(\alpha)$). The shaded area represents the negative welfare cost of inefficient moral hazard (V^{MHI}).

The analysis is different for those living hand-to-mouth. In their case, there is a difference between the marginal consumption-utility cost of medical spending in the uninsured case and the efficient case with insurance. To see this, we recall the expected utility for the hand-to-mouth individuals from Equation (1), which setting $\pi = 1$ we can write as

$$EU_{HTM} = \left[(N-K)u\left(\frac{y-\alpha m}{N}\right) + Ku\left(\frac{y-\alpha m}{N} - \frac{(1-\alpha)m(\alpha)}{K}\right) + h(m(\alpha))\right]. \tag{11}$$

The first order condition for medical spending when uninsured is:

$$FOC_{HTM}^{0} : u'\left(\frac{y}{N} - \frac{m}{K}\right) = h'(m). \tag{12}$$

The difference from the perfect-liquidity case in Equation (9) is that those living hand-to-mouth must absorb the entire medical spending in the shorter time of K periods rather than spreading it across N consumption periods.

The first order condition for the efficient spending level for the hand-to-mouth individuals is

$$FOC^{E}_{HTM}:$$
$$\underbrace{(1-\alpha)u'\left(\frac{y-\alpha m}{N}-\frac{(1-\alpha)m}{K}\right)}_{\substack{\text{Marginal consumption-utility cost of medical spending borne through out-of-pocket costs}}}+\underbrace{\alpha\left(\frac{K}{N}u'\left(\frac{y-\alpha m}{N}-\frac{(1-\alpha)m}{K}\right)+\frac{N-K}{N}u'\left(\frac{y-\alpha m}{N}\right)\right)}_{\substack{\text{Marginal consumption-utility cost of medical spending born through premiums}}}=h'(m). \tag{13}$$

For hand-to-mouth individuals, the marginal consumption-utility costs of medical spending are lower with insurance, resulting in the efficient level of spending being higher than uninsured spending. The key reason is that the insured portion of spending, αm, is spread out across all N consumption periods.[16]

At the same time, there will still be some level of inefficient moral hazard in the hand-to-mouth case. When choosing actual medical spending, the individual takes premiums as given. The first order condition for the chosen level of insurance only includes the first term on the left-hand side of Equation (13) and is given by

$$FOC^{*}_{HTM}:(1-\alpha)u'\left(\frac{y-\alpha m}{N}-\frac{(1-\alpha)m}{K}\right)=h'(m). \tag{14}$$

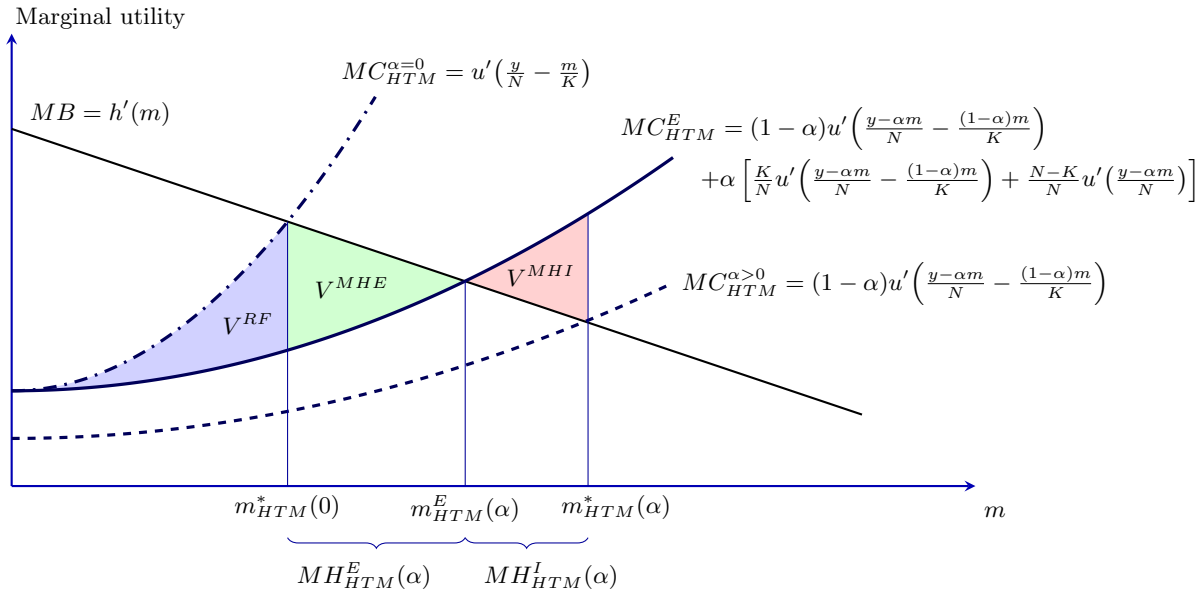
Figure 2 shows the hand-to-mouth case graphically. The full marginal cost under insurance accounting for its impact on premiums (MC^{E}) lies in between the uninsured marginal cost and the marginal cost under insurance. This implies that there is both efficient and inefficient moral hazard.

There are now three shaded regions denoting the welfare impacts of insurance and moral hazard specifically. The V^{MHI} area shaded in pink is the same inefficiency of moral hazard as we saw in the perfect liquidity case – resulting from the fact that the level of spending chosen when insured is higher than the efficient level of spending. However, we also have the green-shaded area V^{MHE}, which represents the utility value of the efficient portion of moral hazard – resulting from the fact that the efficient level of spending is higher than the uninsured level. Moreover, we see the blue-shaded area V^{RF}, that is the risk and financing value. This represents the welfare improvement holding fixed spending at the uninsured level $m^{*}_{HTM}(0)$ that comes from being able to insure against that spending. It is positive (even though $\pi=1$) because of the financing benefit of insurance for the liquidity constrained. Each dollar of spending has a lower impact on consumption utility when it is spread out evenly as insurance premiums across N consumption periods.

---

[16] As such, the marginal consumption-utility cost in the period experiencing the loss is lower with insurance. In addition, the $(N-K)$ consumption periods where no loss is experienced and only premiums are paid have higher levels of consumption and thus lead to a lower average marginal consumption-utility cost for premiums than for uninsured spending. Mathematically we have that: $\frac{K}{N}u'\left(\frac{y-\alpha m}{N}-\frac{(1-\alpha)m}{K}\right)+\frac{N-K}{N}u'\left(\frac{y-\alpha m}{N}\right)<u'\left(\frac{y-\alpha m}{N}-\frac{(1-\alpha)m}{K}\right)<u'\left(\frac{y}{N}-\frac{m}{K}\right)$.

Figure 2 – Analysis of Moral Hazard for Hand-to-Mouth Individuals



Note: Schematic illustration. Assumes hand-to-mouth behavior with $K \leq N$ and sick state happens for certain $\pi = 1$. m is medical spending in the sick state of the world.

This illustration shows a visual example where the net welfare value of moral hazard is positive, since the welfare gain from efficient moral hazard is larger than the welfare loss from inefficient moral hazard. This will not always be the case for hand-to-mouth individuals, but there will always be an efficient component of moral hazard that helps to offset the inefficient moral hazard.

An intuitive explanation of these results is as follows. For someone who has perfect liquidity, optimization implies that they are indifferent between a dollar more of medical spending and a dollar more of consumption on the margin. Thus, when they become insured and face a lower marginal price, the benefit of medical spending is of lower value than an additional dollar of consumption. The analysis is different for the liquidity constrained. When uninsured, they are indifferent on the margin between a dollar more of medical spending and a dollar of consumption in those periods when they pay for their medical care. However, forgoing a dollar of premiums spread out over all consumption periods is not as painful as forgoing that dollar of consumption in those few periods. Thus, an additional dollar of medical spending will be more valuable than the cost of financing that spending with a dollar of premiums. Thus, some of the increased medical spending has higher value to the individual than the full cost to finance it via premiums. It is this fact, that insurance premiums have financing built in, that leads to some of the observed moral hazard being efficient.

3.1.2 Illustrating the Income Effect Using the Full-insurance Case ($\alpha = 1$)

When the sick state is not sure to happen ($\pi < 1$), there is an additional benefit of insurance that we refer to as the "income effect," following Nyman (1999a). The income effect creates efficient moral hazard for both the hand-to-mouth and perfect-liquidity types. The key idea is that every

dollar of insured spending only ends up affecting consumption by its expected costs π via premiums that are shared across those in the sick state (π share) and the non-sick states ($(1-\pi)$ share). We provide formal propositions for the general case in Section 3.3, but here we illustrate the intuition under full insurance.

At full insurance, the expected utility coincides for both types. Because premiums are paid smoothly over all N periods and there are no uninsured cost-shocks, the hand-to-mouth individuals in our model become essentially perfectly liquid. That is, we have

$$EU_{PL}(\alpha=1) = EU_{HTM}(\alpha=1) = \pi\left[Nu\left(\frac{y-\pi m(1)}{N}\right)+h(m(1))\right]+(1-\pi)Nu\left(\frac{y-\pi m(1)}{N}\right). \quad (15)$$

The efficient level of spending $m^E(1)$, which will be the same for both types, will be governed by the first order condition

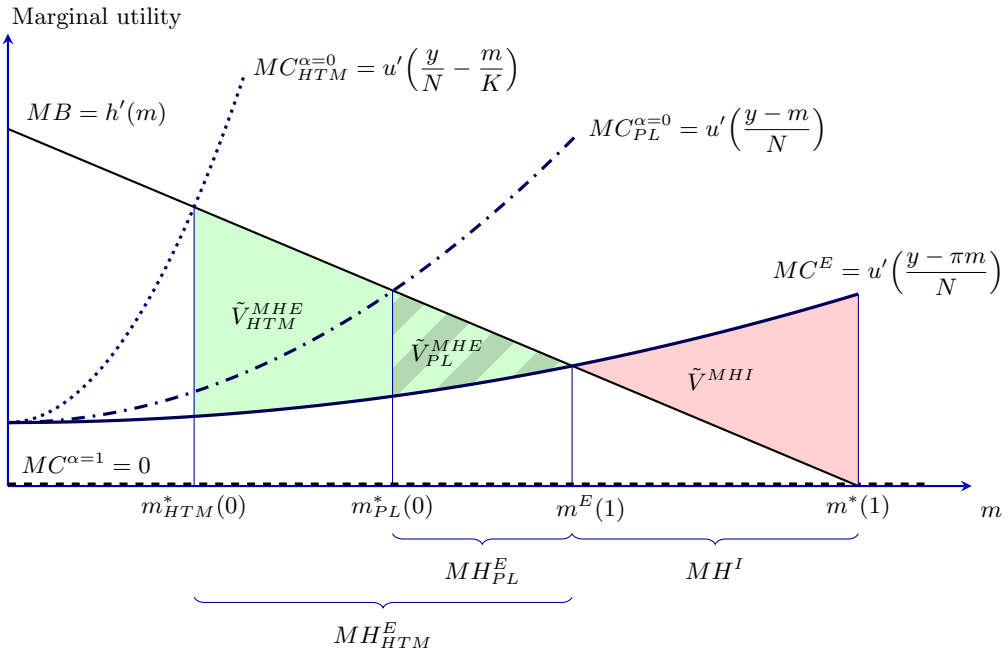
$$FOC^E : u'\left(\frac{y-\pi m}{N}\right) = h'(m). \quad (16)$$

Full insurance guarantees that consumption is steady and based on income minus premiums at $\frac{y-\pi m}{N}$ in every period, regardless of whether the loss occurs or not. As such, the efficient spending level is determined by where the marginal consumption utility from paying premiums equals the marginal health benefit. This level of marginal consumption-utility cost is lower than the uninsured levels.

Figure 3 shows the full insurance case graphically. The income effect for the perfect-liquidity case is generated by the difference between the uninsured marginal consumption-utility cost $u'\left(\frac{y-m}{N}\right)$ (Equation (9)) and the efficient marginal cost, $u'\left(\frac{y-\pi m}{N}\right)$. This difference comes from sharing the medical spending costs with uninsured states through premiums. The gap is bigger when the probability of the sick state π is smaller. The efficient level of spending at full insurance $m^E(1)$ is higher than the uninsured spending, implying that there is efficient moral hazard for the perfect-liquidity case. The positive welfare value of the efficient moral hazard, conditional on the sick state occurring, is given by the darker green shaded region $\tilde{V}_{PL}^{MHE}$. At the same time, because the individual perceives no cost to medical spending under full insurance, they over-consume at the point $m^*(1)$, leading to inefficient moral hazard. In this graphical representation, the net welfare value of moral hazard in the perfect-liquidity case is negative since the losses from inefficient moral hazard are larger than the gains from efficient moral hazard. There can be situations, however, where the efficiency gains from the income effect lead to positive net moral hazard value for those with perfect liquidity. We give conditions for that to happen in the general case in our formal propositions below and provide numerical examples in the next subsection.

The moral hazard benefits of insurance are larger for the hand-to-mouth individuals because they have both the liquidity effect and the income effect of insurance. The liquidity effect can be seen in the figure by comparing the lines for uninsured marginal cost for the hand-to-mouth $u'\left(\frac{y}{N}-\frac{m}{K}\right)$ to the lower uninsured marginal cost under perfect liquidity. The liquidity effect spreads the medical spending across consumption periods, while the income effect spreads the costs across sick and non-sick states via premiums. The level of efficient moral hazard is greater for the hand-to-mouth case because the uninsured level of spending is lower, and the corresponding size of the welfare

Figure 3 – Analysis of Moral Hazard for Both Types under Full Insurance



Note: Schematic illustration. Assumes full insurance $\alpha = 1$. m is medical spending in the sick state of the world. Note that the subscript PL denotes the perfect-liquidity type with $K = N$ and subscript HTM is the hand-to-mouth type with $K < N$. Curves, spending levels, and moral hazard ranges without subscript are the same for both the PL and HTM. The shaded values for the efficient and inefficient moral hazard regions are the values conditional on the sick state being realized. The ex-ante utility values of efficient and inefficient moral hazard as defined in Equations 6 and 7 are given by: $V^{MHE}_{HTM}=\pi\tilde{V}^{MHE}_{HTM}$, $V^{MHE}_{PL}=\pi\tilde{V}^{MHE}_{PL}$, and $V^{MHI}=\pi\tilde{V}^{MHI}$. Note that $\tilde{V}^{MHE}_{HTM}$ is the sum of both green regions.

gain from efficient moral hazard V^{MHE}_{HTM} (denoted by the sum of both green regions) is larger. The level of inefficient moral hazard is the same for hand-to-mouth types as it is for those with perfect liquidity because the chosen level of spending is the same under full insurance. As we show in the formal propositions, because the hand-to-mouth types have both the liquidity effect and income effect, the net value of moral hazard is more likely to be positive for the hand-to-mouth types. An implication is that a benevolent social planner would typically set a higher level of insurance for hand-to-mouth types than for perfect-liquidity types.

3.1.3 The Efficient Level of Spending

Given information on the level of insured and uninsured spending, our model shows how to infer the efficient level of medical spending. Understanding the efficient level of spending has practical implications for the design of health care systems. In addition to optimizing demand-side cost-sharing parameters, policymakers heavily rely on supply-side restrictions—such as prior authorization, coverage determinations, utilization management, and global budgets—to control spending. However, to design effective supply-side restrictions, the insurer or policymaker requires a target: what is the efficient level of medical utilization?

For this illustration, we focus on the efficient level of spending at full insurance ($\alpha = 1$).[17] Recall that $m^E(\alpha)$ is the level of spending that maximizes the individual's utility, subject to the constraint that medical spending is paid for via premiums. That is, $m^E(1) \equiv \arg\max_m \; Nu(\frac{y-\pi m}{N}) + \pi h(m)$. The efficient level of spending satisfies the first order condition $u'(\frac{y-\pi m^E(1)}{N}) = h'(m^E(1))$, as seen in the curve MC^E in Figure 3.

Note that at full insurance, $m^E(1)$ will be the same for both individuals with perfect liquidity and with liquidity constraints. What will differ, however, is the relationship between $m^E(1)$ and the uninsured spending of the two types. As seen in Figure 3, $m^E(1)$ will be much higher relative to $m^*_{HTM}(0)$ than it is relative to $m^*_{PL}(0)$.

With an additional parametric assumption, we can use data on observed spending when uninsured ($m^*(0)$) and insured ($m^*(1)$) to infer m^E. Specifically, we make the assumption that $h(m)$ is quadratic such that

$$h(m) = \frac{1}{2}a_1 m^2 + a_2 m + a_3. \tag{17}$$

This form fulfills all previous assumptions we made on $h(m)$ and can flexibly approximate curvature in the relevant region of health spending. It also implies the marginal product of health spending declines linearly.

Using the two observed spending amounts and the assumption that spending decisions are optimal we can estimate parameters a_1 and a_2 of the health benefit function. Note that here and in all other applications of our model in this paper, the parameter a_3 is irrelevant. This lets us derive the following expression that defines the efficient level of medical spending for each type i:

$$\frac{u'(c_i^E(1))}{u'(\tilde{c}_i(0))} = \frac{m^*(1) - m_i^E(1)}{m^*(1) - m^*(0)} \equiv 1 - s_i^E(0,1) \tag{18}$$

where $c_i^E(1) = \frac{y-\pi m_i^E(1)}{N}$ is consumption when fully insured and $\tilde{c}_i(0) = \frac{y}{N} - \frac{m^*(0)}{K_i}$ is the consumption in the periods when the loss is borne for type i. The left-hand side is the ratio of the marginal utility cost of healthcare spending in two different cases: being paid for via premiums at the efficient level, over being paid for out-of-pocket when uninsured. The right hand side of the equation is that share of observed moral hazard that is inefficient. The share of observed moral hazard that is *efficient*, when moving from uninsurance to full insurance, is denoted $s_i^E(0,1)$. It allows for an intuitive expression of the efficient level of spending as $m_i^E(1) = m_i^*(0) + s_i^E(0,1) \times [m_i^*(1) - m_i^*(0)]$.

With perfect liquidity and $\pi = 1$, one can see that the efficient level of spending is the uninsured level: $c^E(1) = \tilde{c}_{PL}(0)$. However, when $\pi < 1$, $m^E_{PL} > m^*_{PL}(0)$, as even for people with perfect liquidity, income is transferred to the sick state from the healthy state, and marginal utility of consumption for spending paid for via premiums is lower than marginal utility of spending when uninsured. For hand-to-mouth individuals facing high marginal utility cost when sick (since medical spending is spread across fewer periods), the difference in marginal utilities can be large, meaning

[17] A more general result is provided in Section 3.3

efficient spending is much closer to insured spending than to uninsured spending. This quantifies the intuition from Figure 3.

3.2 Numerical examples

This subsection provides numerical examples to help quantify the plausible size of the welfare effects of moral hazard and to show that the effects of liquidity constraints are not simply an alternative way of capturing the effects of a more concave utility function.

We use a set of linear approximations to the marginal consumption-utility costs in the first-order conditions that govern uninsured, efficient, and chosen levels of spending. We use first-order Taylor expansions of the marginal consumption utility terms around the marginal utility at baseline consumption $u'\left(\frac{y}{N}\right)$. We then normalize the entire first order condition by dividing by $u'\left(\frac{y}{N}\right)$. This results in a set of linear marginal (consumption-utility) cost curves that are in monetary units with intuitive functional forms. Appendix C provides the full set of derivations for general levels of insurance and probability of sickness.

The linear approximations to the first-order condition governing uninsured spending is given by:

$$1 + r\frac{m_i(0)}{K_i} \approx H'(m_i(0)) \tag{19}$$

The left-hand side is the marginal consumption-utility cost of medical spending. The slope of the marginal costs is governed in part by r, which is the Arrow-Pratt measure of absolute risk aversion at baseline consumption ($r = -\frac{u''(y/N)}{u'(y/N)}$). Without insurance, the marginal cost of medical spending grows with m at a rate of $\frac{r}{N}$ in the perfect-liquidity case where $K_i = N$ and at a rate of r in the $K_i = 1$ hand-to-mouth case. The marginal consumption-utility cost is N times steeper for the hand-to-mouth individuals who have to absorb the entire spending in one period.

Next, examine the first-order condition for efficient spending under full insurance. As we highlighted in the prior subsection, under full insurance, this condition is the same for the perfect-liquidity and hand-to-mouth cases, so we can suppress the type subscript on $m^E(1)$. The first order condition for efficient spending in the linear approximation is:

$$1 + \pi r\frac{m^E(1)}{N} \approx H'(m^E(1)) \tag{20}$$

The efficient marginal cost under full insurance rises with m at a rate of $\pi\frac{r}{N}$. The income effect of insurance generates efficient moral hazard for the perfect-liquidity case by lowering the slope of the marginal cost relative to the uninsured case by a factor of π. For the hand-to-mouth case, the efficiency of moral hazard under full insurance arises because the marginal cost under full insurance is reduced by both π and $\frac{1}{N}$.

For our numerical example we pin down these approximate marginal cost curves by assuming $r = 0.0006$. This is the level of absolute risk aversion that corresponds to a Constant Relative Risk

Aversion utility function $u = \frac{c^{1-\rho}}{1-\rho}$ with $\rho = 3$ and evaluated at baseline consumption for annual income $y = \$60,000$ and monthly consumption periods $N = 12$.

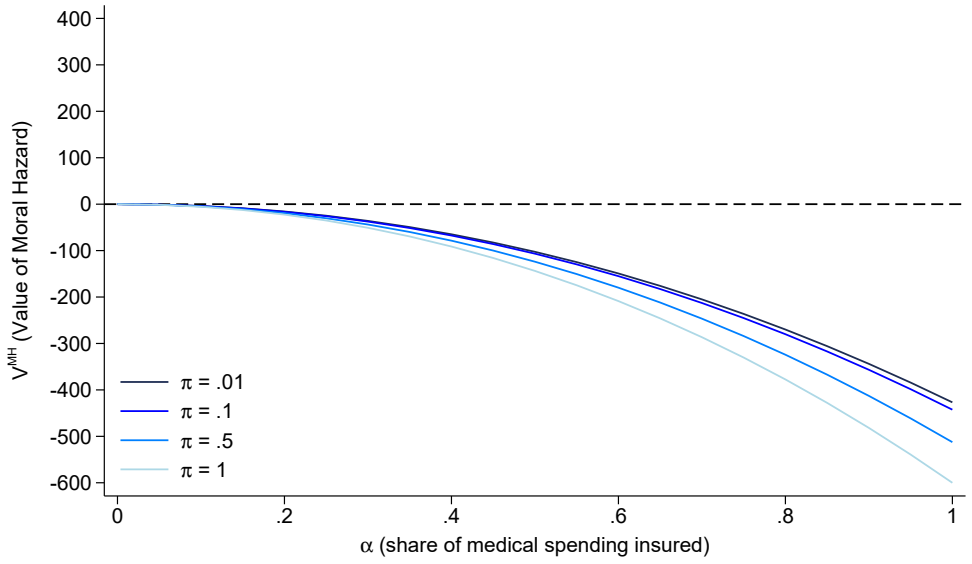
The H' terms on the right-hand side of the first-order conditions are the marginal health benefit of medical spending (h') normalized by dividing by $u'\left(\frac{y}{N}\right)$. We assume that the marginal health benefit is linear, which is equivalent to assuming the quadratic form for h from Equation (17). Under this linearity assumption, the H' curve can be pinned down if we know two levels of spending. For our numerical example, we assume that uninsured spending for those with perfect liquidity is $m^*_{PL}(0) = \$3,000$ and that the spending by those individuals under full insurance would be $m^*(1) = \$4,000$. With all of the curves pinned down, the welfare values of moral hazard illustrated in the shaded regions of Figures 1 to 3 become triangles that can be easily quantified.

Figure 4 plots the resulting total moral hazard value ($V_i^{MH} = V_i^{MHE} + V_i^{MHI}$) for insurance levels ranging from no ($\alpha = 0$) to full insurance ($\alpha = 1$) and for four different levels of the probability of sick state ranging from a 1 percent chance to a sure sick state. Panel (a) plots the results for individuals with perfect liquidity. We see that for this numerical example, the total moral hazard value is negative for those with perfect liquidity and worse as the level of insurance gets higher. At higher levels of insurance, moral hazard reduces welfare by around \$300 to \$600. Due to the income effect there is some level of efficient moral hazard when $\pi < 1$, which can be seen by noting that the total value of moral hazard at every level of insurance is higher for lower values of π. However, the value of efficient moral hazard is low relative to the losses from inefficient moral hazard, leading to negative moral hazard value across the range of values explored here.

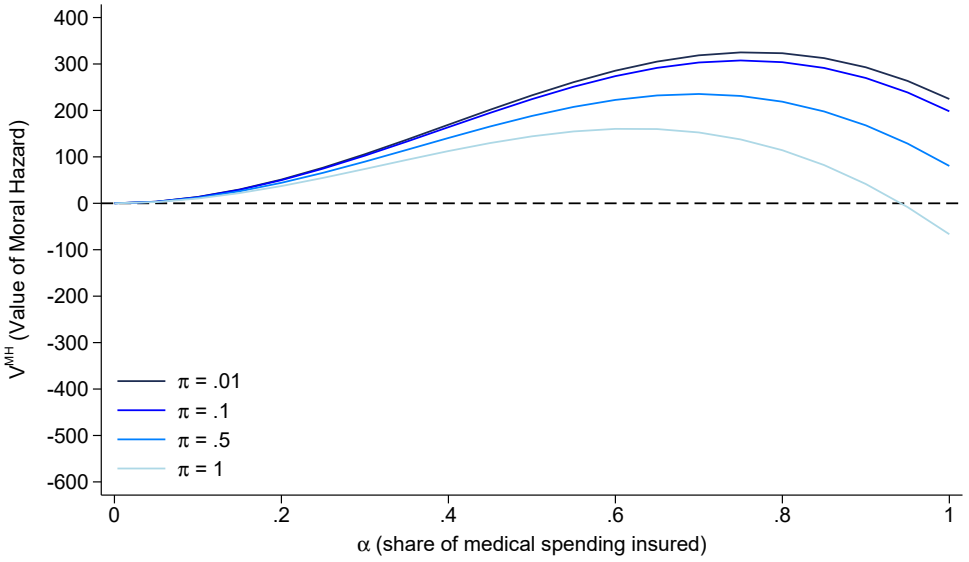
For hand-to-mouth individuals, however, the value of moral hazard is often significantly positive as can be seen in panel (b). In this example, the value of moral hazard peaks at relatively high levels of insurance, around 80% insurance for probabilities of sickness of 10% or less. Noting that the risk and financing value is rising with the level of insurance, this implies that the optimal level of insurance will be close to full insurance for the hand-to-mouth. The value of moral hazard can be quite significant, reaching over \$300 for higher levels of insurance against lower probabilities of sickness. We also note that even when the sick state is sure to arise, the net value of moral hazard is positive up to nearly full insurance (crossing to be negative around 95% insurance). This means that in this example, hand-to-mouth individuals derive positive moral hazard benefits from insurance that helps them finance medical expenditures they are sure to need and that it can be beneficial to finance nearly all such expenditures. This is true even despite the fact that insurance generates some degree of inefficient spending due to distortions in the perceived price of care.

It might be tempting to think about liquidity constraints as simply representing a case of a very concave utility function, but there is an important distinction between the moral hazard value of insurance under liquidity constraints and under strong utility curvature but perfect liquidity. The efficiency gains of moral hazard under insurance arise in the hand-to-mouth case because the marginal consumption-utility cost curve is very steep when uninsured relative to when insured. It is also true that a much stronger level of utility curvature can generate a similarly steep marginal cost curve when uninsured for those who have perfect liquidity. This case would arise for those

Figure 4 – Numerical Examples of Moral Hazard Value



(a) Perfect Liquidity $K_i = N$

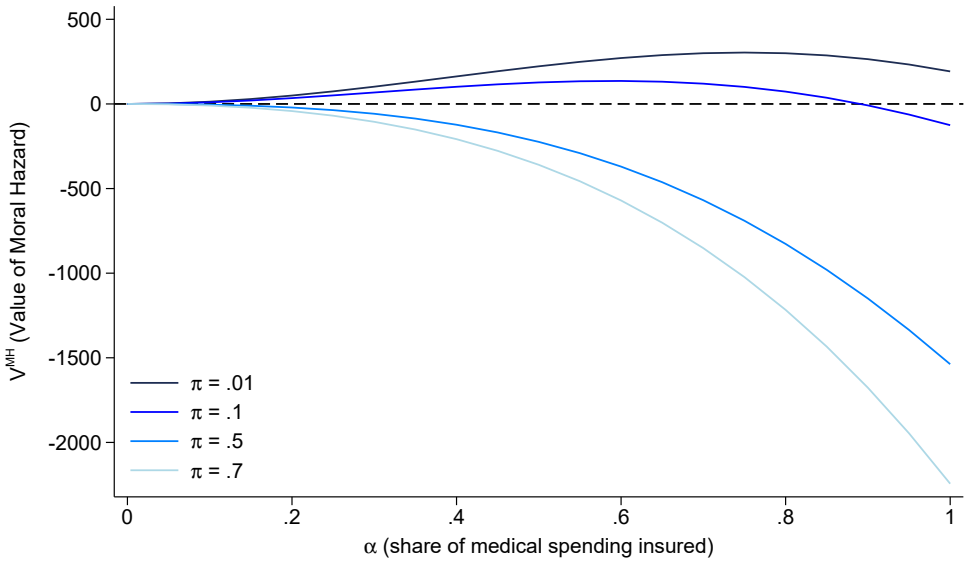


(b) Hand-to-Mouth $K_i = 1$

Note: These figures show the net value of moral hazard for different levels of insurance (α) and probability of sick state (π). We use the linear approximation to the marginal utility of consumption outlined in this subsection. The parameters governing the numerical exercise are that a) the number of consumption periods $N = 12$, b) the Arrow-Pratt measure of absolute risk aversion $r = 0.0006$, c) the uninsured spending under perfect liquidity is $m^*_{PL}(0) = \$3,000$, and d) the maximum spending observed under full insurance is $m^*(1) = \$4,000$. As described in the text, we assume a linear marginal utility of health function $h'(m) = c - bm$ and the parameters are pinned down by the assumptions about $m^*_{PL}(0)$ and $m^*(1)$. Panel a) shows the perfect liquidity case where $K_i = N$ and Panel b) shows the full hand-to-mouth case where $K_i = 1$.

with low annual incomes or when the cost of medical care is high, such that medical spending is traded off against essential consumption. However, the efficiency benefits without liquidity constraints arise only from the income effect, which depends on a low probability of the sick state. For those with perfect liquidity, as the probability of the sick state approaches 1, the uninsured spending approaches the efficient level of spending. However, for those who are hand-to-mouth, the uninsured spending stays below the efficient level even for $\pi = 1$.

Figure 5 – Numerical Example Perfect Liquidity and Strong Utility Curvature



Note: This figure shows the net value of moral hazard for different levels of insurance (α) and probability of sick state (π) for the perfect-liquidity case where $K_i = N$. We use the linear approximation to the marginal utility of consumption outlined in this subsection. The parameters governing the numerical exercise are that a) the number of consumption periods $N = 12$, b) the Arrow-Pratt measure of absolute risk aversion $r = 0.0072$, c) the uninsured spending under perfect liquidity is $m^*_{PL}(0) = \$3,000$, and d) the maximum spending observed under full insurance is $m^*(1) = \$4,000$. As described in the text, we assume a linear marginal utility of health function $h'(m) = c - bm$ and the parameters are pinned down by the assumptions about $m^*_{PL}(0)$ and $m^*(1)$.

Figure 5 provides an example of this. For this figure we repeat the numerical analysis from Figure 4 for the perfect liquidity case but under the assumption that absolute risk aversion is 12 times the original absolute risk aversion in the prior figure ($r' = 12r = 0.0072$). As such, in this example the uninsured marginal consumption-utility cost for the perfect liquidity individual is now the same as the uninsured marginal consumption-utility cost for the hand-to-mouth in the numerical example in Figure 4 panel (b). Because this marginal consumption-utility cost curve when uninsured has the same slope, with the observed spending assumptions ($m^*(0) = \$3,000$, $m^*(1) = \$4,000$), we also have the same calibrated marginal utility of health function $H'(m)$ from the prior numerical example for the hand-to-mouth case. What differs now is that the gap between the uninsured and efficient marginal cost curves depends only on the probability of loss (i.e., the income effect) and not on $K < N$. We see now that the net value of moral hazard is positive at low levels of

π. In particular, if the likelihood of the sick state were 1%, the highly-risk-averse individual with perfect liquidity derives up to around $300 in value from moral hazard at moderately high level of insurance and significantly positive value from moral hazard even under full insurance. However, as the likelihood of the sick state rises, the value of moral hazard falls dramatically and becomes substantially negative. Those with high coefficient of absolute risk aversion benefit from moral hazard when premium costs are low, but as the likelihood of sickness rises and premiums go up, the impact of moral hazard is even worse than it would be if the individual had a less concave utility function (compare Figure 5 to Figure 4 panel (a)). In contrast, when the high sensitivity to uninsured spending costs comes from the lack of liquidity, as in Figure 4 panel (b), the moral hazard value of insurance comes primarily from its liquidity benefit and is much less sensitive to the probability of the sick state.

3.3 Formal propositions

In this section, we develop a set of formal propositions that describe how liquidity constraints affect individuals' behavior and value of insurance. All proofs are provided in the appendix.

First, we lay out common results for perfect-liquidity and hand-to-mouth individuals. Let $i \in \{PL, HTM\}$ denote type of liquidity, with K_i denoting the number of consumption periods over which medical payments are effectively spread, so $K_{PL} = N$ and $K_{HTM} < N$. Define the following condition, which relates the difference in marginal utility in the sick and healthy states to the curvature of utility:

$$(N - 2K_i\pi)u'\left(\frac{y}{N} - \frac{m_i^*(0)}{K_i}\right) - 2(N - K_i\pi)u'\left(\frac{y}{N}\right) > m_i^*(0)\frac{N - K_i\pi}{K_i}u''\left(\frac{y}{N} - \frac{m_i^*(0)}{K_i}\right) \quad (21)$$

We have a set of results that apply to both types of individuals when $\pi < 1$:

Proposition 1. *For $\alpha \in (0, 1]$ and $\pi \in (0, 1)$, and for both types PL and HTM the following statements hold:*

1. $m_i^*(\alpha) > m_i^E(\alpha)$.
2. $u''' > 0$ *is a sufficient condition for* $m_i^E(\alpha) > m_i^*(0)$ *and hence* $V_i^{MHE}(\alpha) > 0$.
3. *If Equation* (21) *holds for type i, then there exists* $\hat{\alpha}_i > 0$ *such that* $V_i^{MH}(\alpha) > 0 \quad \forall \alpha \in (0, \hat{\alpha}_i]$
4. *If Equation* (21) *holds for type i, then there exists* $\hat{\alpha}_i' > \hat{\alpha}_i$ *such that* $V_i(\alpha) > 0 \quad \forall \alpha \in (0, \hat{\alpha}_i']$.

Part 1 confirms the standard intuition that individuals with insurance choose a higher level of medical spending than is efficient. Part 2 covers a different result, known from Nyman (1999a): the efficient level of medical spending is larger than the level that would be chosen if the individual were uninsured. Here, we assume $u''' \geq 0$ to obtain global monotonicity of $m_i^E(\alpha)$ over the full

coverage range. Note that it is only a sufficient but not a necessary condition.[18] Further, for the most important case of full insurance, which is also the case discussed by Nyman, concavity of u is sufficient to show that $m_i^E(1) > m_i^*(0)$. Part 3 lets us state the argument of Nyman (1999a) formally: there can be situations in which insurance has a positive moral hazard value. The inequality in equation (21) shows a condition sufficient to yield a positive moral hazard value. Note that the condition is more likely to be satisfied if π gets smaller. This reflects the argument in Nyman (1999a) that smaller probabilities of getting sick lead to higher transfers of healthy people to sick people through the insurance system. Finally, Part 4 shows that if that same condition, there is a range of insurance levels where the total value of insurance is positive.

Next, we show how results are qualitatively different between individuals with perfect liquidity and liquidity constrained types when $\pi = 1$, and establishes the results discussed in Section 3.1.1:

Proposition 2. *Let $\pi = 1$. Then, for $K_{HTM} < N$, Proposition 1 statements 1-4 also hold. However, for individuals with perfect liquidity ($K_{PL} = N$) only statement 1 holds. Further, for all $\alpha \in (0, 1]$, $m_{PL}^E(\alpha) = m_{PL}^*(0)$, $V_{PL}^{MHE}(\alpha) = 0$, $V_{PL}^{MH}(\alpha) < 0$, and $V_{PL}(\alpha) < 0$.*

The proposition establishes a novel implication of liquidity constraints: there is efficient moral hazard even if there is no risk, as the socially optimally level of spending is higher than the level that would be chosen when uninsured even when $\pi = 1$. Moreover, it shows a sufficient condition for the entire moral hazard value of insurance to be positive when $\pi = 1$. Since this result does not rely on $\pi < 1$, it is independent of that shown by Nyman (1999a).

In what way does living hand-to-mouth change the value of insurance for individuals? Proposition 2 answers this question for the case of certain losses. When there is no uncertainty, under perfect liquidity, insurance only decreases the welfare of the individual. For hand-to-mouth individuals, on the other hand, the value of insurance can indeed be positive.

When introducing risk, the comparison gets more complicated. There is now the potential of efficient moral hazard for individuals with perfect liquidity due to the effect described by Nyman (1999a). This effect gets stronger with lower loss probabilities and higher medical spending.

We can nevertheless make certain comparisons. Proposition 3 quickly results from the different ways in which individuals can finance their medical spending. At any level of positive cost-sharing (including being fully uninsured), hand-to-mouth individuals will choose less medical spending than perfectly liquid individuals:

Proposition 3. *For $K_{HTM} < N$ and $K_{PL} = N$, $m_{HTM}^*(\alpha) < m_{PL}^*(\alpha)$ for all $\alpha \in [0, 1)$.*

At full insurance, the experience of hand-to-mouth and perfectly liquid individuals is the same – smooth premium payments and no out-of-pocket medical costs. Thus, the chosen level of medical spending will be the same for the two types ($m_{HTM}^*(1) = m_{PL}^*(1)$), as well the efficient level of medical spending ($m_{HTM}^E(1) = m_{PL}^E(1)$). Then, the first part of Proposition 4 follows from the

[18] Also, most commonly used utility functions, including CARA and CRRA, satisfy this condition (Brockett and Golden, 1987).

definitions of V^{MH} and V^{MHE}, and says that for the hand-to-mouth individual, the value of both moral hazard and efficient moral hazard, will be higher than for the perfectly liquid individual.

Proposition 4. *For $K_{HTM} < N$ and $K_{PL} = N$ and $\alpha = 1$, it holds that*

1. *$V^{MHE}_{HTM} > V^{MHE}_{PL}$ and $V^{MH}_{HTM} > V^{MH}_{PL}$, and*
2. *$s^E_{HTM}(0,1) > s^E_{PL}(0,1)$ where $s^E_i(0,\alpha) = \frac{m^E_i(\alpha) - m^*_i(0)}{m^*_i(\alpha) - m^*_i(0)}$.*

The second part of the proposition defines $s^E_i(0,\alpha)$ as the share of the moral hazard response to coverage level α that is efficient. Under full insurance, this share, which can be connected to the value of insurance as we show in Proposition 8 below, is strictly higher for hand-to-mouth individuals than for people with perfect liquidity.

While Proposition 4 shows that, at full insurance, the value of moral hazard is higher for hand-to-mouth individuals, that value may still be negative. The next proposition considers the range of insurance values for which the value of moral hazard is positive. After making a simplifying assumption on the health benefits function, Proposition 5 shows that hand-to-mouth individuals have a larger range of insurance values for which moral hazard value is positive, than do perfectly liquid individuals.

Proposition 5. *For $K_{HTM} < N$ and $K_{PL} = N$, assume $u''' \geq 0$, and that benefits from medical spending are approximately linear in the relevant ranges of m_i. If for both types there exists an $\hat{\alpha}_i \in (0,1)$ such that $V^{MH}_i(\alpha) > 0$ if and only if $\alpha \in (0, \hat{\alpha}]$, then $\hat{\alpha}_{HTM} > \hat{\alpha}_{PL}$.*

This result also links our discussion of efficient and inefficient moral hazard to the value of insurance. We divide the value of insurance according to $V_i(\alpha) = V^{MH}_i(\alpha) + V^{RF}_i(\alpha)$ and know that $V^{RF}_i(\alpha)$ is always positive. The value of insurance thus hinges on the value of moral hazard not being too negative. Because $V^{MH}_i(\alpha) = V^{MHI}_i(\alpha) + V^{MHE}_i(\alpha)$, we know that $V^{MH}_i(\alpha)$ can only be positive if the value of efficient moral hazard is larger than the value of inefficient moral hazard. By showing that, conditional on its existence, the threshold $\hat{\alpha}$ below which $V^{MH}_i(\alpha) > 0$ is larger for hand-to-mouth individuals than for perfect liquidity individuals, we demonstrate the influence of liquidity constraints on how moral hazard interacts with the value of insurance.

Finally, we consider the optimal level of insurance α^*_i that an individual would choose for themselves, given that they faced actuarially fair premiums. We define this optimal level as $\alpha^*_i = \arg\max_{\alpha \in [0,1]} EU_i(\alpha)$ subject to the constraint that premiums $p = \alpha \pi m^*_i(\alpha)$. This is also the level of insurance that a social planner, with the goal to maximize individual welfare, would choose.

Using the same simplifying assumption on the health benefit function as above, Proposition 6 shows that the optimal level of insurance coverage is higher for hand-to-mouth individuals than for those with perfect liquidity. We discuss the implications for this result in the next section.

Proposition 6. *For $K_{HTM} < N$ and $K_{PL} = N$, if benefits from medical spending are approximately linear in the relevant ranges of m_i and the optimal insurance coverage is characterized by an interior solution, then $\alpha^*_{HTM} > \alpha^*_{PL}$.*

Our final formal propositions consider the share of moral hazard that is efficient. They show how it can be calculated from data and how it relates to the value of moral hazard, and hence, to the value of insurance as a whole. Proposition 7 shows how to infer the share of moral hazard that is efficient from chosen spending levels under two levels of insurance. It is a general version of the statement previously provided in Equation (18) for moving from uninsured to full insurance. Note that in this general form, the efficient share of moral hazard can be negative in case medical spending at α_1 is already larger than efficient medical spending at α_2. An example for such a situation is given in Appendix E.2.

Proposition 7. *Assume that the benefit of medical spending is quadratic as described in Equation* (17) *and that we observe medical spending for two coverage levels* $m^*(\alpha_1)$ *and* $m^*(\alpha_2)$ *with* $\alpha_2 > \alpha_1$. *Define* $c(\alpha, K) = \frac{y-\pi\alpha m^*(\alpha)}{N} - \frac{(1-\alpha)m^*(\alpha)}{K}$ *and* $E[u'_E(\alpha, K)] = \alpha\frac{N-\pi K}{N}u'\left(\frac{y-\pi\alpha m_K^E(\alpha)}{N}\right) + \left(\frac{\alpha\pi K}{N} + (1-\alpha)\right)u'\left(\frac{y-\pi\alpha m_K^E(\alpha)}{N} - \frac{(1-\alpha)m_K^E(\alpha)}{K}\right)$. *The share of moral hazard that is efficient when moving from* α_1 *to* α_2 *is implicitly defined by*

$$\frac{E[u'_E(\alpha_2, K)] - (1-\alpha_1)u'(c(\alpha_1, K))}{(1-\alpha_2)u'(c(\alpha_2, K)) - (1-\alpha_1)u'(c(\alpha_1, K))} = \frac{m_K^E(\alpha_2) - m^*(\alpha_1)}{m^*(\alpha_2) - m^*(\alpha_1)} \equiv s^E(\alpha_1, \alpha_2). \tag{22}$$

Finally, Proposition 8 connects the share of moral hazard that is efficient to the value of insurance. A challenge is that this value is determined by the marginal utilities, while the magnitude of a behavioral response to a price change is typically given by the second derivative of the utility function (Mas-Colell et al., 1995). We make a pair of simplifying assumptions to yield a simple expression for this result. First, we assume that the health benefit function is quadratic, consistent with our assumptions in Equation (17) of Section 3.1.3. Second, we assume that we can approximate the utility function with a second-order Taylor approximation, a common assumption in empirical applications (e.g. Barseghyan et al., 2013).[19]

Recall that for any level of insurance coverage, we have defined the share of efficient moral hazard as $s^E(0, \alpha) = \frac{m^E(\alpha)-m^*(0)}{m^*(\alpha)-m^*(0)}$ above. Using this definition and the two simplifying assumptions, the following result decomposes the value of insurance into expressions for the value of moral hazard and the value of risk and financing.

Proposition 8. *Assume that benefits from medical spending are quadratic and that the consumption utility function can be approximated by a Taylor approximation of the second order. Let* r *be the coefficient of absolute risk aversion at* $\frac{y}{N}$. *The value of insurance in monetary units can be written*

[19] It also is equivalent to the assumption of CARA utility and normally distributed losses as used in an application by Azevedo and Gottlieb (2017) and in our own Section 5.

as $\frac{V(\alpha)}{u'(y/N)} = \frac{V^{MH}(\alpha)}{u'(y/N)} + \frac{V^{RF}(\alpha)}{u'(y/N)}$ *where*

$$\frac{V^{RF}(\alpha)}{u'(y/N)} = \frac{1}{2}\pi(2-\alpha)\alpha(m^*(0))^2\left(\frac{1}{K}-\frac{\pi}{N}\right)r \tag{23}$$

$$\frac{V^{MH}(\alpha)}{u'(y/N)} = \frac{m^*(\alpha)-m^*(0)}{m^*(0)}\left[s^E(0,\alpha)\frac{V^{RF}(\alpha)}{u'(y/N)} - \left(1-s^E(0,\alpha)\right)\frac{1}{2}\pi\alpha m^*(0)\left(1+\frac{\pi}{N}rm^*(\alpha)\right)\right] \tag{24}$$

Proposition 8 provides a tractable, "dollar-valued" decomposition of the value of insurance by normalizing by marginal utility and shows the role of curvature through the local coefficient of absolute risk aversion $r = -\frac{u''(y/N)}{u'(y/N)}$.

The risk-and-financing component $V^{RF}(\alpha)$ shows how this value scales with the size of uninsured medical spending $m^*(0)$, curvature r, the final term $\left(\frac{1}{K}-\frac{\pi}{N}\right)$, which compares the per-period burden of out-of-pocket payments when they must be financed over K periods to the per-period burden of insurance premiums smoothed over N periods.

The moral hazard component $V^{MH}(\alpha)$ depends on the decomposition of observed moral hazard into its efficient and inefficient share. The expression for $V^{MH}(\alpha)$ makes it explicit that the moral hazard value depends positively on the efficient share $s^E(0,\alpha)$. In particular, V^{MH} loads on V^{RF} through the term $s^E(0,\alpha)$. When more of the additional spending is efficient, the same consumption-smoothing forces that generate V^{RF} also increase the welfare gain from the behavioral response.

3.4 Interpreting the Coverage Elasticity of Demand for Medical Care

The elasticity of medical spending in insurance coverage is an empirical object commonly used to link theory to data (Zeckhauser, 1970; Besley, 1988; Finkelstein et al., 2019).[20] In this section, we show how liquidity constraints changes the interpretation of this elasticity.

To keep the exposition simple, we abbreviate consumption when premiums are paid $\frac{y-\alpha\pi m}{N}$ as $A(m)$, and consumption when medical expenses are also paid as $B_i(m) = \frac{y-\alpha\pi m}{N} - \frac{(1-\alpha)m}{K_i}$. We defer formal arguments to Appendix D.

We denote $e_i^\alpha = \frac{\alpha}{m_i^*(\alpha)}\frac{\partial m_i^*(\alpha)}{\partial \alpha}$ as the insurance coverage elasticity of the demand for medical care for type i. Because we have normalized the price of medical care to 1 for uninsured individuals, the price is $1-\alpha$ when people have insurance. e_i^α is thus an inverse version of the price-elasticity of demand for medical care.

We begin with individuals with perfect liquidity. Combining the definition of the elasticity with first-order condition for optimal insurance coverage, and following Besley (1988), we show in

[20] The elasticity approach is also very common in unemployment insurance (Baily, 1978; Chetty, 2006, 2008; Hendren, 2017).

Appendix D that the following condition must hold at the optimal insurance coverage α^*:

$$1 + e_{PL}^{\alpha^*} = \frac{u'(B_{PL}(m_{PL}^*(\alpha)))}{\pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A(m_{PL}^*(\alpha)))}. \tag{25}$$

We can interpret Equation (25) in the same terms as the Baily-Chetty equation is interpreted for unemployment insurance (Chetty, 2006). $e_{PL}^{\alpha^*}$ is the markup people are willing to pay for funds in the sick state compared to receiving them in their *ex-ante* overall expected financial situation. The difference to the unemployment insurance situation is that medical expenses in our model are financed through premiums and not taxes, so it is inappropriate to label the elasticity as a fiscal externality. More appropriately one could label it as the efficiency loss due to moral hazard.

When individuals are hand-to-mouth and $K < N$, the relationship looks slightly different. Following the same steps as above, we have

$$1 + e_{HTM}^{\alpha^*} = \frac{u'(B_{HTM}(m_{HTM}^*(\alpha)))}{\pi\frac{K}{N} u'(B_{HTM}(m_{HTM}^*(\alpha))) + (1-\pi\frac{K}{N})u'(A(m_{HTM}^*(\alpha)))}. \tag{26}$$

The equation implies that $e_{HTM}^{\alpha^*}$ is the markup people are willing to pay for funds in the sick *and liquidity constrained* state compared to receiving them in their *ex-ante* overall expected financial situation. Appendix D shows that, as a result, hand-to-mouth individuals are willing to pay a larger markup at any given coinsurance level below full coverage. This changes the interpretation of an observed elasticity – for any observed demand elasticity, the optimal level of coinsurance is higher. This mirrors our result from Proposition 6.

4 Revising Estimates of the Value of Medicaid

4.1 Model

We now bring our model of liquidity constraints to the Oregon Health Insurance Experiment (OHIE), and show how liquidity constraints affect the interpretation of moral hazard and the welfare consequences of the Medicaid expansion. We adapt the framework of Finkelstein et al. (2019, hereafter FHL). We use the same data as FHL to calibrate our model and estimate the value of a year of Medicaid to recipients, focusing specifically on how assumptions about liquidity affect individuals' implied valuation of empirical moral hazard.

We adapt the optimization-based approaches laid out by FHL. Our model and FHL's share common assumptions: people select a level of health spending given a health shock by optimally trading off medical spending and consumption under the assumption that their utility is separable between health and consumption. FHL discuss limitations of this approach: e.g. individuals may make mistakes in allocating health spending, or observed health spending may represent corner solutions such that the first order conditions do not hold. These limitations apply to our analysis as well. Neither our results below nor the original FHL analysis using this approach are perfect

measures of the value of Medicaid to recipients. Our goal, however, is to examine how the estimates would change if Medicaid recipients lived hand-to-mouth rather than having perfect liquidity.

We make a set of modeling decisions. First, we adapt the setting to the binary loss framework we have already laid out. While, in principle, we could model a continuous distribution of losses, doing so substantially complicates the assumptions needed to go from the data to the model. Second, charity care plays an important role for these individuals: the population gaining access to Medicaid did not pay for the full cost of their health care when uninsured. We model charity-care as a lump-sum transfer ψ which uninsured individuals get in case of sickness that takes the form of deduction from the final health-care bill.[21]

We use a willingness to pay approach to describe the value of insurance, following FHL. Equations (3) through (5) define the value of insurance and its components in utility terms. To convert that to a monetary evaluation, we define $W_i(\alpha, m, p)$ as the expected utility of type-i individuals with coverage level α, medical spending m, and premium p. Consistent with the procedure by FHL, we can then implicitly define the value of Medicaid, γ_i, as the premium recipients would pay that would give them the same level of utility as being uninsured:

$$W_i(1, m^*(1), \gamma_i) - W_i(0, m^*(0), 0) = 0. \tag{27}$$

As before, we can split the value of insurance into a sum of a transfer term, a risk and financing term, and a moral hazard term: $\gamma_i = \gamma_i^T + \gamma_i^{RF} + \gamma_i^{MH}$. Here, γ_i^T is the amount of money transferred due to insurance coverage for the expected medical spending individuals would consume even in the absence of any insurance coverage. It is given by

$$\gamma_i^T = \pi(m^*(0) - \psi). \tag{28}$$

The risk and financing component, γ_i^{RF}, is implicitly defined by WTP under insurance but holding fixed spending at the uninsured level:

$$W_i(1, m^*(0), \gamma_i^{RF} + \gamma_i^T) - W_i(0, m^*(0), 0) = 0. \tag{29}$$

Finally, the willingness to pay for the moral hazard component, γ_i^{MH}, is the remainder of value.

The risk and financing motive can be further decomposed into WTP for risk reduction, γ_i^R, and WTP for financing across consumption periods, $\gamma_i^F \equiv \gamma_i^{RF} - \gamma_i^R$. To value the risk reduction apart from the financing benefit of insurance, we introduce a hypothetical insurance contract for which all premiums have to be paid in the K consumption periods that also feature losses. This isolates

[21] FHL take a different approach and model charity care as lowering the marginal price of medical spending for the uninsured. In that model, uninsured individuals only pay a percentage of the costs of care. We lay out such a model in Appendix E.2, and show that it implies that uninsured individuals were over-consuming medical care, relative to the efficient benchmark. Because we deem this unrealistic, we focus on the lump-sum model here.

the WTP for risk reduction. The following equation implicitly defines γ_i^R:

$$(N - K_i)u\left(\frac{y}{N}\right) + K_i u\left(\frac{y}{N} - \frac{\gamma_i^R + \gamma_i^T}{K_i}\right) + \pi h(m^*(0)) - W_i(0, m^*(0), 0) = 0. \tag{30}$$

Note that if $K = N$, the definition in Equation (30) reduces to Equation (29) implying $\gamma_{PL}^{RF} = \gamma_{PL}^R$ and $\gamma_{PL}^F = 0$. This reflects the fact that perfectly liquid individuals have no need for financing medical spending across consumption periods.

Finally, we follow FHL and make a set of functional form assumptions: utility of consumption has the constant relative risk aversion form, with $u(c) = \frac{c^{1-\sigma}}{1-\sigma}$ and $\sigma = 3$.

For the health production function, we follow the literature on numerical calibrations of moral hazard models (Einav et al., 2013; Azevedo and Gottlieb, 2017; Marone and Sabety, 2022) and use a quadratic function of the form $h_i(m) = \frac{1}{2}a_{1,i}m^2 + a_{2,i}m + a_{3,i}$ from Equation 17.[22] Because medical spending is chosen optimally, the first-order conditions and observed spending amounts when fully insured and uninsured identify the parameters of $h_i(m)$.[23] When fully insured by Medicaid, individuals face zero marginal cost of spending and choose $m^*(1)$ to maximize $h(m)$, implying $a_{2,i} = -a_{1,i}m^*(1)$. For the uninsured, we assume charity care operates as a lump-sum deduction from the medical bill, with additional care purchased at full marginal cost. From the first-order equation that determines the choice of $m_i^*(0)$, we get:

$$a_{1,i} = \frac{u'\left(\frac{y}{N} - \frac{m^*(0)-\psi}{K_i}\right)}{m^*(0) - m^*(1)}. \tag{31}$$

4.2 Data

We follow FHL in most assumptions on the data and diverge from them only to accommodate our two-state model.[24] FHL use two approaches for determining consumption and we follow their simpler approach, which they call the "consumption proxy approach". Here, they start with an average value of consumption for the low-income uninsured from the Consumer Expenditure Survey (CEX), which they estimate at $9,214 for the time period and sample they use. They then add to that the average per-capita out-of-pocket spending for the uninsured in the OHIE to get a total baseline average income level of $y = $9,505. For $N = 12$ consumption periods, this translates to an income of $792 per consumption period in our model. One other difference in our approach is how to accommodate household size. FHL in their baseline analysis assume that reported medical costs are shared among household members. All of our results instead assume the individual in

[22] Note that the analyses in the references use this function for continuous losses which changes the form slightly; see also Section 5.2. The function used here is formally equivalent for a single sick state.

[23] As before, $a_{3,i}$ is irrelevant for the value of insurance or choice of medical spending.

[24] We used the replication package available through the *Journal of Political Economy* to stay as close to the original assumptions as possible.

question bears both the full benefits and the full costs of Medicaid. This allows us to avoid difficult decisions about how to account for children in household risk sharing.

The OHIE data contains records of medical spending for both the insured (treatment compliers) and the uninsured (control compliers). We categorize a state of "sickness" as a state of the world in which an insured individual has non-zero medical spending. Following FHL, we use the procedure of Abadie (2003) and calculate this probability for treatment compliers, arriving at $\pi = 79.21\%$. Using the same data and approach, we also calculate the expected spending of control and treatment compliers. When dividing by π, this renders $m^*(0) = \$3,404$ and $m^*(1) = \$4,554$. FHL use the OHIE data to determine that charity care covers 79% of expenses for the uninsured. We thus set $\psi = 0.79 \times \$3,404$. All resulting parameters are summarized in Table 1 below.

Table 1 – Parameters used to determine the value of Medicaid

Variable	Value
y, annual income	\$9505
π, probability of loss	79.21%
$m^*(0)$, uninsured medical spending in case of sickness	\$3404
$m^*(1)$, insured medical spending in case of sickness	\$4554
ψ, charity-care transfer for the uninsured in the sick state	\$2689
$\pi(m^*(1) - m^*(0))$, moral hazard effect (in terms of expected value)	\$911
$\pi(m^*(1) - \psi)$, net cost of providing Medicaid	\$1477

Note: Source: Authors' calculations from FHL data. y is calculated for the applicable population in the CEX. π, $m^*(0)$, and $m^*(1)$ are causal estimates from the OHIE data. ψ is calculated as $0.79 \times m^*(0)$.

4.3 Estimates

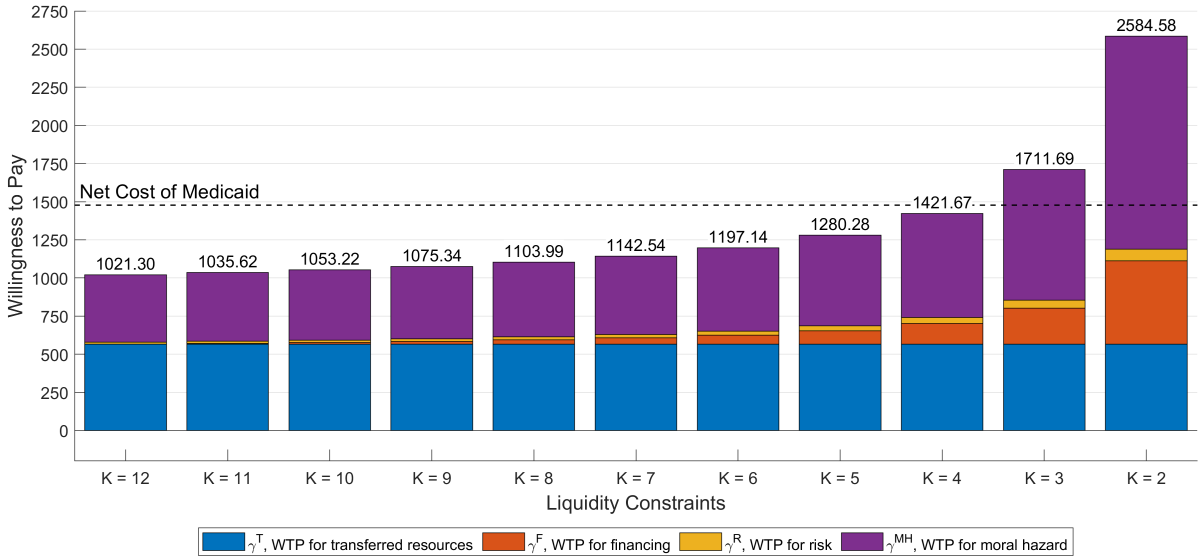
Our main results on how liquidity constraints affect the value of Medicaid are shown in Figure 6.[25] The graph displays the total value of Medicaid for different assumptions on liquidity constraints, ranging from the case of perfect liquidity ($K = N$) to the case of $K = 2$, which allows individuals to spread expenses across two months. This is close to minimal liquidity. We do not consider the case of $K = 1$ in the main text because it creates levels of consumption in the uninsured case when the sick state occurs that fall below the per-period version of the consumption floor imposed by FHL.[26]

Under perfect liquidity, we estimate a willingness-to-pay for Medicaid of \$1,021. This is somewhat smaller than the estimate provided by FHL. The difference arises because our binary loss distribution leads to less extreme outcomes than the continuous distribution considered by FHL. This is reflected in the fact that the willingness to pay for risk reduction is remarkably small at

[25] The results in this and the following graph are shown in table form in Table E1 of the Online Appendix.

[26] We provide results for the $K = 1$ cases under three different assumptions (lower consumption floor, higher income, and lower uninsured spending) in Appendix E.3. In all three cases, the estimated value of Medicaid increases strongly compared to the results discussed here.

Figure 6 – Impact of Liquidity Constraints on Value of Medicaid



Note: Authors' calculations from FHL data. All values are in \$US and correspond to the values for γ_i and its components defined in Equations (27) through (30). Parameters are set as in Table 1 and follow the approach of FHL. We assume that consumption utility has a CRRA utility function with a coefficient of relative risk aversion equal 3. Health production is quadratic with the parameters calibrated from first-order conditions. N is set to 12.

only about \$13; though, recall that these individuals receive charity care covering 79% of their uninsured spending, so the potential loss this term is paying to insure is only about \$715 (or 7.5% or their income). What is reassuring is the result on γ^{MH}. Access to Medicaid led to an average increase in expected medical spending of \$911 compared to being uninsured. With perfect liquidity, individuals value this increased spending at \$442 or at roughly 49% of its cost – the standard case of wasteful moral hazard. The ratio of value to cost in our results is similar to that in FHL: In their consumption-based optimization approach the ratio of value to cost of medical care at uninsured prices is mechanically equal to 50% thus showing that our results are consistent with theirs.[27]

Estimated willingness to pay for Medicaid increases as K decreases, and does so convexly. We focus on the results for the lowest reported level of liquidity ($K = 2$). The total willingness to pay for individuals with low liquidity is more than two and a half times as large as for individuals with perfect liquidity. As explained above, this is because Medicaid provides funds in consumption periods when marginal utility is very high and consumers value this transfer accordingly.

[27] To see this, consider that their Transfer Term after implementation of their Assumption 4 on their Equation (14) is (in their notation) equal $\rho(0)\frac{1}{2}(E[m(1;\theta)] + E[m(0;\theta)])$. Splitting this into $\gamma^T + \gamma^{MH}$ leads to the equivalent term $\rho(0)E[m(0,\theta)] + \rho(0)\frac{1}{2}(E[m(1,\theta)] - E[m(0,\theta)])$. Thus, for them $\frac{\gamma^{MH}}{\rho(0)(E[m(1,\theta)]-E[m(0,\theta)])} = \frac{1}{2}$.

Examining the decomposition, the transferred resources component is constant across all levels of K. All other components increase with as K decreases. At $K = 2$, there is a substantial willingness to pay from the financing motive which is now equal to \$545. The reason for this strong financing value of insurance is that with $K = 2$, the uninsured out-of-pocket medical spending after charity care of \$715 is absorbed in only two periods, equating to 45% of income for those consumption periods. For the perfect liquidity case, spreading this same out-of-pocket cost across $K = 12$ periods meant only a 7.5% hit to each period's income. Insurance allows liquidity constrained individuals to finance the spending across consumption periods, which they value highly. The willingness to pay from the risk motive is also increased; though at \$78 it is still relatively small.

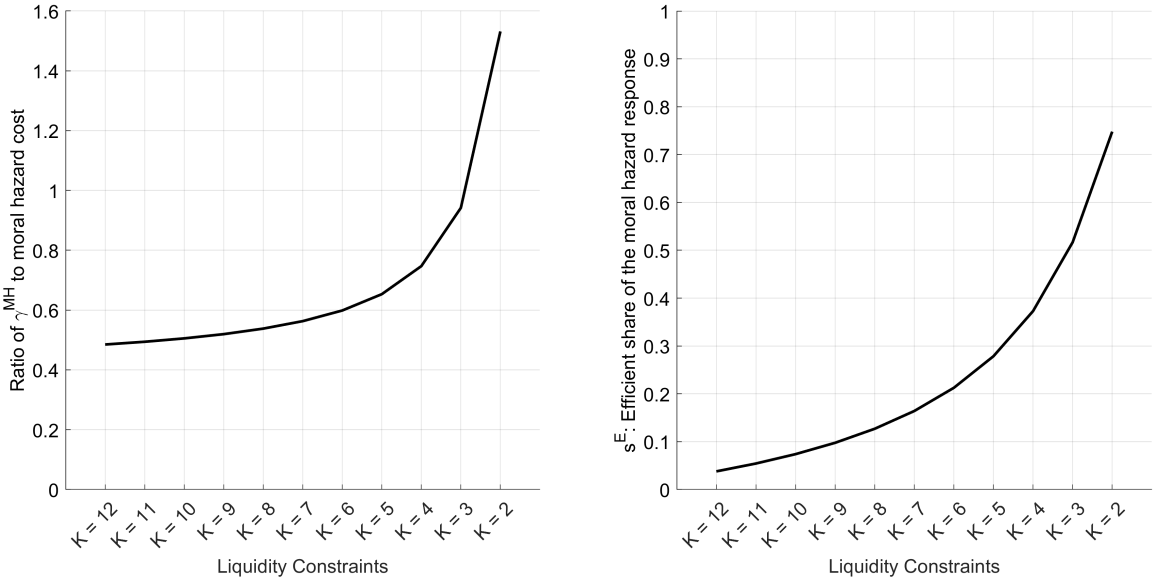
The largest driver of the increased willingness to pay for Medicaid is the valuation of moral hazard which amounts to \$1,396 for $K = 2$. This value is more than three times higher than in the perfect liquidity case. This is a reflection of our earlier argument: with hand-to-mouth individuals, moral hazard does not have to be an inefficient adjustment, but can rather reflect the possibility of financing health care costs across more consumption periods than the one in which the costs appear.

Panel (a) of Figure 7 shows the ratio of γ^{MH} to the cost of additional medical spending under Medicaid. For perfectly liquid individuals, the ratio is 0.49. It rises to around 1 for $K = 3$ and over 1.5 for $K = 2$. This demonstrates that the liquidity-constrained are willing to pay more (via premiums) for the additional medical spending than the cost to provide it, while individuals with perfect liquidity have the "standard" result of willingness to pay less than cost. The moral hazard component is inefficient if Medicaid recipients can spread medical expenditures across time, but if they must absorb them in a limited number of periods, moral hazard is net efficient.

Next, we examine the efficient level of spending m^E for individuals who have received Medicaid. This provides further context for γ^{MH} and is useful to Medicaid programs considering supply side constraints on care. Recall that $m^E(\alpha)$ is the level of spending that maximizes the individual's utility, subject to the constraint that medical spending is paid for via premiums. When uninsured, individuals receive a charity care transfer in case of sickness. We incorporate the expected value of this transfer ($\pi\psi = \$2,130$) into the initial wealth. Individuals then have to pay actuarially fair premiums for Medicaid under efficient spending levels. In panel (b) of Figure 7, we display s^E: the share of observed moral hazard that is efficient. For perfectly liquid individuals, it is relatively low at about 4%. However, as individuals become more liquidity constrained, the efficient share increases. At $K = 2$, 75% of the increased medical spending due to insurance coverage is efficient.

Figures 6 and 7 show that under liquidity constraints, the value of Medicaid increases. It is interesting to contrast the calculated values of the willingness to pay with the cost of providing Medicaid. Aside from the administrative cost about which we can make no credible statement, we can calculate the cost of providing medical care under Medicaid. Table 1 informs us that under full coverage, individuals spend \$4,554 on medical care when sick, which equates to expected payments of \$3,607. In the absence of insurance, individuals consume medical spending of \$3,404, of which 79% are provided by charity care. This equates to expected cost of the charity care transfer of

Figure 7 – Evaluation of the Moral Hazard Reaction



(a) Ratio of willingness to pay for moral hazard and its cost

(b) Share of efficient moral hazard

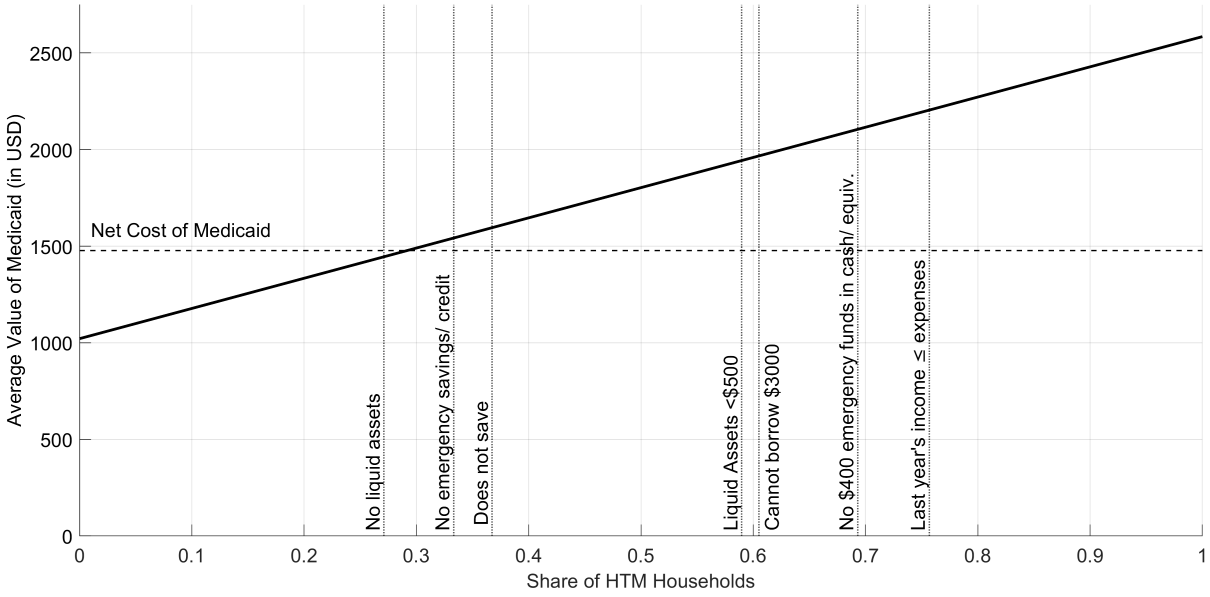
Note: Source: Authors' calculations from FHL data. Panel (a) shows γ^{MH} divided by its cost. Cost $\pi(m^*(1) - m^*(0))$ is \$911. Panel (b) shows $s^E(0,1) = \frac{m^E(1)-m^*(0)}{m^*(1)-m^*(0)}$ with $m^E(1)$ defined implicitly through $h'(m^E(1)) - u'((y + \pi\psi - \pi m^E(1))/N) = 0$.

\$2,130. The net cost of providing Medicaid to an individual is thus \$1,477. The willingness-to-pay of perfectly liquid individuals in this binary model is below net costs. However, for people with liquidity constraints, the valuation of Medicaid exceeds the net cost of providing it for values of K around 3 or 4. For individuals with $K = 2$, willingness-to-pay exceeds net costs by more than \$1,000.

How liquidity-constrained is this population? Evidence on liquidity-sensitivity of medical utilization among low-income seniors (e.g. Gross et al., 2022) suggests that these constraints do in fact matter. While we lack data on the full distribution of liquidity constraints and borrowing costs that the Medicaid population faces, we can provide a range of estimates. Figure 8 illustrates how the average value of Medicaid changes depending on the assumed fraction of Medicaid recipients that are liquidity constrained (with $K = 2$) versus perfectly liquid. The horizontal line indicates the net cost of providing Medicaid. Using data from the Survey of Consumer Finances and the Survey of Household Economics and Decision-Making, we estimate what fraction of the population in the U.S. below 138% of the federal poverty level (a population similar to those eligible for Medicaid) satisfies various conditions that proxy for being liquidity constrained. The vertical lines indicate the fraction of individuals meeting criteria such as having zero liquid assets (27%), having less than \$500 in liquid assets (59%), or lacking \$400 in emergency savings in cash or cash equivalents (69%). Appendix F provides more detail on these measures. Note as another point of comparison that

Aguiar et al. (2025) classify about 40% of the PSID (across all income levels) as hand-to-mouth, and that Appendix F shows that low-income individuals are much more likely to display indicators of hand-to-mouth status.

Figure 8 – Value of Medicaid by Share of Hand-to-Mouth Households



Note: The figure displays the value of γ, the estimated willingness to pay for Medicaid, as it is reported in the total values of Figure 6 for possible shares of liquidity constrained ($K = 2$) households in the population. All other households are assumed to have perfect liquidity ($K = 12$). The vertical lines show the values of potential indicators for this share from the Survey of Consumer Finances and the Survey of Household Economics and Decision-making as they are defined in Appendix F. The dashed horizontal line shows the net cost of providing Medicaid.

The valuation of Medicaid depends heavily on the fraction of individuals subject to liquidity constraints. For instance, if individuals with zero liquid assets are assumed to be hand-to-mouth (and all others have perfect liquidity), Medicaid's average value is about \$1440, whereas assuming those without \$400 in emergency funds behave as hand-to-mouth yields an average valuation of roughly \$2100. The variation, and the result that the interpretation of moral hazard's welfare impact changes under the hand-to-mouth assumption suggests there would be substantial policy value to additional data on the recipients' of different government insurance programs liquidity constraints and the underlying patterns of medical spending and consumption so that policy recommendations can reflect the value of insurance under liquidity constraints.

Finally, note that our results will depend not only on households' liquidity conditions but also on the distribution of out-of-pocket medical expenditure across time for the low-income uninsured. We do not know this distribution, but it is likely to be more concentrated in time than the spending of the insured, as uninsured people will screen out lower value care and spend more on urgent health needs. We use the 2018 Truven Marketscan sample of 24-64 year olds continuously enrolled in

insurance (as described in Ericson and Sydnor, forthcoming). In this sample, the average fraction of total expenses in the highest expenditure month is 0.54. For smaller expenses of under $1000, this value is higher at about 0.67.[28] Given these numbers for wealthier people, we think that assuming hand-to-mouth consumption and the concentration of medical payments in two consumption periods for a portion of the population is a reasonable approximation to the distribution of out-of-pocket medical expenditures of the uninsured.

5 Selection on Moral Hazard and the Value of Vertical Choice

5.1 Introduction

Our numerical simulations as well as the application to Medicaid show that liquidity constraints can significantly increase the value of insurance coverage. In this section, we demonstrate that this is not the only consequence of including liquidity constraints when modeling insurance demand. Heterogeneous liquidity constraints can also impact the welfare consequences of vertical choice in insurance menus.

Recent research on the evaluation of vertical choice analyzes the issue using a model of selection on moral hazard (Marone and Sabety, 2022). This concept, introduced by Einav et al. (2013), describes a setting in which individuals have private information not only about their health status, but also about their behavioral response to coverage—that is, the elasticity of their medical spending with respect to the out-of-pocket price. The private information about this responsiveness effectively adds a second dimension of adverse selection in addition to selection on health status. However, while variation in moral hazard preferences may indeed be quite important, liquidity constraints can be another source of background heterogeneity that is linked to utilization responses to insurance coverage.

In this section, we show that heterogeneity in moral hazard can arise from differences in households' liquidity constraints rather than from heterogeneous preferences for medical-care utilization. Assuming perfect liquidity can lead to biases in the estimation of the selection on moral hazard model parameters. Moreover, liquidity constraints generate different out-of-sample predictions for willingness to pay for insurance and thus different welfare conclusions about the impact of insurance menu design. Estimation strategies that ignore liquidity constraints and instead treat utilization differences as preference-driven moral hazard might therefore misdiagnose the source of heterogeneity and potentially make incorrect policy conclusions about the value of vertical choice.

We use the Azevedo and Gottlieb (2017) adaptation of the Einav et al. (2013) model, and follow their derivations and use their parameter estimates. We conduct simulations from a population that has heterogeneity in liquidity constraints, but not in their moral hazard preference parameter. We

[28] Actual OOP expenses are likely more concentrated overtime (e.g. once an individual reaches the maximum out-of-pocket amount, there will be no more cost-sharing), but that will depend on the insurance plan an individual is in. We do not attempt to model those insurance plans because what matters for our sample is the structure of charity care provided for the uninsured, about which there is little data.

then ask: what would an economist conclude about the optimal single plan to offer the population if they estimated a selection on moral hazard model rather than a liquidity constraints model on these data? While (incorrectly) inferred variation in moral hazard preference parameters can explain much of the variation in the data, the conclusions about which plan to offer are different. Using the data-generating model with variation in liquidity constraints, the welfare-maximizing single plan covers 75 percent of medical costs. In contrast, using the estimated selection on moral hazard model, the recommended plan would cover only 38 percent of medical costs.

We then explore the implications for insurance menu design. Previous work has argued that the gains from vertical choice of insurance levels is small since selection can undermine any gains to choice that result from preference heterogeneity (Ericson and Sydnor, 2017; Marone and Sabety, 2022). However, this work has not examined the impact of variation in liquidity constraints. When we allow for selection on moral hazard but impose perfect liquidity, we find negligible to negative values of having choice. However, once we allow for variation in liquidity constraints, there are large welfare gains to offering choice in insurance plans.[29]

The remainder of the section proceeds as follows. We first layout the model. We then simulate choices from a model of a population that has variation in liquidity constraints but not the moral hazard preference parameter. We show what parameters would result from the misspecified model that assumes no liquidity constraints and loads heterogeneity on the moral hazard preference parameter. We then show what insurance plan design would maximize welfare in the liquidity-constrained data-generating model versus the misspecified selection-on-moral-hazard model. Finally, we solve for the equilibrium with vertical choice in both the misspecified model and the data generating model with liquidity constraints and show the value of vertical choice.

5.2 Model

Our model adapts the Einav et al. (2013) and Azevedo and Gottlieb (2017) model to allow for N consumption periods and liquidity constraints in which medical expenses must be borne in $K \leq N$ periods. Individuals (indexed i) suffer a continuously distributed loss l and in reaction choose how much to spend on medical care m. Losses shift the optimal level of medical spending, as the benefits to medical spending are $h_i(m,l) = (m-l) - \frac{1}{2H_i}(m-l)^2$. Benefits to medical spending are measured in monetary units and accrue in all periods.[30] Here, H_i is the moral hazard parameter that dictates the responsiveness of medical care demand to insurance coverage.

---

[29] If instead we consider a world in which there is heterogeneity in the preference parameters governing medical care utilization and that people were liquidity constrained (e.g., $K = 1$ or $K = 3$) but not heterogeneously so, we estimate that similar to the original Marone and Sabety (2022) results, there is little value to vertical choice. However, there is a significant difference in what the optimal single contract would be under vertical choice for low K values than for perfect liquidity $K = N$.

[30] One way to rationalize this is to think of the health shock in the first period as persistent and so any medical spending to address it provides benefits in all periods. However, the model is also similar to one where benefits of medical spending only appear the first period and are scaled by factor $1/N$.

The insurance contract provides straight coinsurance, with coverage level α at premium p. The individual has income y/N in each period. Given this set up, the utility of an individual with K_i conditional on health shock l and medical care m is given by:

$$u_i(m, l; \alpha, p) = K_i v_i \left(\frac{y - p + h_i(m, l)}{N} - \frac{(1-\alpha)m}{K_i} \right) + (N - K_i) v_i \left(\frac{y - p + h_i(m, l)}{N} \right). \quad (32)$$

The function $v_i(\cdot)$ is a CARA utility function where r_i is the coefficient of absolute risk aversion.

To determine $m_i^*(l; \alpha)$, the optimal choice of m for a realized loss l in a contract with coverage α, we take the first order condition for m in the above equation. This equation does not have an analytical solution for $m_i^*(l; \alpha)$, and so will generally need to be solved numerically.

However, in the special case of perfect liquidity where $K = N$, the individual's utility is

$$u_i(m, l; \alpha, p) = N v_i \left(\frac{y - p + h_i(m, l) - (1-\alpha)m}{N} \right). \quad (33)$$

In this case, there is a simple analytical form for the privately optimal consumption of health care: $m_i^*(l; \alpha, K_i = N) = l + H_i \cdot \alpha$. Here, H_i can be interpreted as the increase in utilization that would occur when an individual went from no insurance ($\alpha = 0$) to full insurance ($\alpha = 1$).

Working under the assumption of perfect liquidity, Einav et al. (2013) show that the moral hazard parameter can be identified if you observe an individual's expected medical spending in two different contracts. Given that the analytic form for optimal medical care is linear in α with slope H_i in the absence of liquidity constraints, then an estimator of H_i is

$$\widehat{H}_{i|K=N} = \frac{\overline{m}_i(\alpha_1) - \overline{m}_i(\alpha_2)}{\alpha_1 - \alpha_2}. \quad (34)$$

However, this is not a valid estimator for H_i if $K_i < N$. In this case, a numerical solution based on the implicit definition of $m_i^*(l; \alpha)$ and multiple draws from the loss distribution is necessary to determine $\widehat{H}_{i|K<N}$.

5.3 Simulation and Estimation

We simulated 1,000 individuals with heterogeneous liquidity constraints, parameterized by K_i. Each individual faces a normal distribution of losses l with mean μ_i and standard deviation σ_i, and has a coefficient of absolute risk aversion r_i. The values of (μ_i, σ_i, r_i) were drawn from a lognormal distribution with means and log-variance-covariance matrix similar to the one used by Azevedo and Gottlieb (2017) and summarized in Appendix Table G1. However, we fix the moral hazard parameter H to be equal to the mean value used by Azevedo and Gottlieb (2017) for all individuals, thus shutting down that preference-based source of heterogeneity in moral hazard. For liquidity constraints, we let K be iid uniformly distributed across the discrete values 1 through 12.

We simulate an insurance plan change as the source of identifying variation that might be used by an econometrician to estimate a model with selection on moral hazard. We assign individuals to

two different insurance plans: first, $\alpha_{pre} = 0.5$ and then $\alpha_{post} = 0.9$. Then, for each individual, we create two simulated data points: $\bar{m}_i^*(\alpha_{pre})$ and $\bar{m}_i^*(\alpha_{post})$, their average medical utilization choice in each contract. Using the average utilization, rather than a single draw of utilization, reduces simulation error in this exercise.

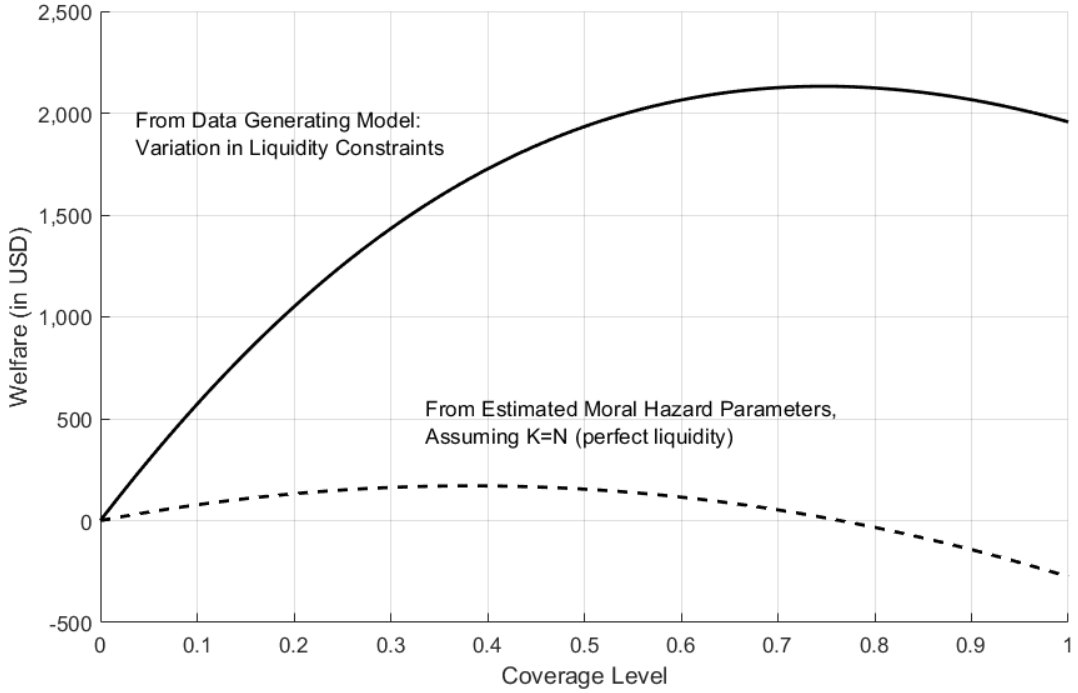
We then take the perspective of the naive econometrician who assumes individuals are not liquidity constrained. For each individual, we use Equation (34) to estimate a moral hazard parameter $\widehat{H_i}$, assuming $K_i = N$. Appendix Figure G1 shows the selection on moral hazard model's estimated $\widehat{H_i}$. The estimator recovers the correct value when $K = N$ and individuals are not liquidity constrained. However, when K is smaller than N (stronger liquidity constraints), the estimated $\widehat{H_i}$ is both higher and more variable between individuals. The former is because with $K_i < N$, the individual is more reactive in their medical spending than would be predicted for perfect liquidity. The misspecified model takes this increased reactivity as evidence for higher levels of H_i. The increased variability in $\widehat{H_i}$ appears due to the interaction of risk aversion and H_i under liquidity constraints.

5.4 Results

Optimal Level of Insurance. The choice between the two models matters for optimal policy, with the selection on moral hazard implying a lower optimal level of insurance. We calculate average welfare if the whole simulated population were in a single plan, for each possible level of insurance. This replicates the decision of an employer choosing which plan to offer, or of a health insurance marketplace designer choosing a single "tier" to offer. For each insurance level α in our grid, we calculate the average willingness-to-pay (WTP) across all individuals for this plan given their parameters $\{\mu_i, \sigma_i, r_i, H_i, K_i\}$, and the expected cost to the insurer. Welfare in plan α is simply average WTP minus average cost. This definition normalizes welfare to zero when $\alpha = 0$. We then repeat this exercise assuming $K_i = N \; \forall \; i$ and using the estimated $\widehat{H_i}$ instead of H_i.

Figure 9 plots the welfare results, for each coinsurance level, if the entire population were enrolled in a plan with that insurance level. The solid line shows welfare from the data-generating model, which includes individuals of varying liquidity constraints K_i. Here, insurance is always welfare increasing relative to being uninsured, and welfare is maximized with a plan that insures 75% of medical spending. In contrast, the dashed line shows the welfare estimates from a model that (incorrectly) assumes $K = 12$ and lets the moral hazard parameter H vary. With these assumptions, welfare is maximized at a coverage level of $\alpha = 38\%$ and is negative for coverage levels above 77%. Further, we observe that the welfare impact of different insurance levels is substantially larger for the liquidity-constrained model.

Figure 9 – Welfare from Insurance Coverage, According to Liquidity Constrained and Selection on Moral Hazard Models



Note: Lines reflect the welfare of putting the entire simulated population in a contract with each coverage level. Solid line: welfare from the true data-generating model, with K_i uniformly distributed between 1 and 12 and H_i common across all individuals. Dashed line: welfare from the model that assumes $K_i = N$ for all i and estimates H_i from Equation (34).

The Value of Vertical Choice. To investigate the implications for the value of vertical choice, we use a simplified version of the research design of Marone and Sabety (2022).[31] In Table 2, we consider possible choice menus of sufficiently different insurance contracts: coverage levels in 10 percentage point increments. For every possible baseline coverage level, we use the results from Figure 9 and report in the "No Choice" columns the welfare from placing the entire population in a single insurance plan, where welfare is again WTP minus cost.

Then, for each baseline coverage level, we calculate a series of equilibria: for each plan with higher coverage (in 10 percentage point increments), we calculate the equilibrium that would result from offering that plan alongside the baseline coverage plan. Marginal premiums are set to the difference in average cost between the plans, and individuals choose optimally between the two options given their estimated preferences. That is, we solve for the Einav et al. (2010) equilibrium and consider the average per capita welfare in the market (see Einav et al., 2013, for a similar

[31] Their analysis differs from ours in that they allow all potential contracts to be traded, not just two contracts. In addition, Marone and Sabety (2022) evaluate vertical choice by solving for the welfare-maximizing premiums, not the competitive equilibrium premiums that we show.

Table 2 – Benefits from Vertical Choice

	Assuming Perfect Liquidity			Data Generating Model		
Baseline Coverage Level	No Choice Mean Welfare (per capita)	Best 2nd Choice Option	Mean Welfare With Choice (per capita)	No Choice Mean Welfare (per capita)	Best 2nd Choice Option	Mean Welfare With Choice (per capita)
0%	\$0.00	60.00%	\$6.16	\$0.00	80.00%	\$1958.58
10%	\$77.63	60.00%	\$82.00	\$571.78	70.00%	\$1998.16
20%	\$131.89	60.00%	\$134.77	\$1049.58	80.00%	\$2041.45
30%	\$162.79	50.00%	\$164.22	\$1433.21	80.00%	\$2093.35
40%	\$170.31	-	-	\$1725.88	80.00%	\$2125.50
50%	\$154.46	-	-	\$1933.69	80.00%	\$2155.48
60%	\$115.24	-	-	\$2064.20	90.00%	\$2159.57
70%	\$52.65	-	-	\$2125.10	90.00%	\$2156.07
80%	\$-33.31	-	-	\$2123.50	90.00%	\$2123.52
90%	\$-142.64	-	-	\$2065.73	-	-

Note: The table shows average per capita welfare when the baseline coverage level is mandatory for all individuals (columns 2 and 5). Columns 3 and 6 give the best second choice option with a coverage level higher than the baseline coverage. Individuals now need to choose between one of the two options. A dash indicates that no 2nd choice option increases welfare. Columns 5 and 7 show the average per capita welfare resulting from the menu with vertical choice. Columns 2 through 4 show values for the model that assume $K_i = N$ for all i and estimates H_i from Equation (34). Columns 5 through 7 show values for the data generating model with variation in K_i.

welfare analysis). We present the coverage level of the best second option to offer alongside the baseline coverage level, and report the mean welfare in the equilibrium that results from offering that choice.

Table 2 provides the results of this simulation for both the estimated selection on moral hazard model that assumes $K_i = N$, as well the data generating model that includes variation in liquidity constraints. For the former model, the table mirrors the results in Marone and Sabety (2022). There is very little gain from vertical choice. Welfare is maximized when only one mandatory contract with 40% coverage is offered and there is no vertical choice. Vertical choice only offers a benefit for coverage levels of 30% or less and even then only provides little benefit, never exceeding \$6.16 per capita.

In contrast, the data generating model implies larger gains from choice. Here vertical choice offers a benefit for every baseline coverage level except for 90%. The welfare maximizing choice is a menu of 60% and 90% coverage. In this situation, vertical choice increases consumer welfare by \$95 *per capita* compared to the situation when only 60% is offered. Compared to the highest level of welfare without vertical choice (70% coverage), there is a welfare increase of \$34 *per capita*. The choice between the two models thus again matters distinctly for optimal policy. The selection on moral hazard model suggests little value for vertical choice. If instead, variation in medical care utilization stems from liquidity constraints, vertical choice is welfare increasing in many situations.

These simulations are conducted under an assumption about the underlying source of heterogeneity that is driving different moral hazard responses from insurance. It would be an interesting extension for future work to explore the types of contract variations that could distinguish liquidity

constraints and selection on moral hazard empirically. One intuition is that variations in contracts that affect the extent to which costs are smoothed over time but do not affect the total marginal price of care should help identify heterogeneity in K independent of that in H.

6 Conclusion

This paper reexamines the interpretation of observed moral hazard from health insurance in the presence of liquidity constraints. We introduce a framework that decomposes moral hazard into efficient and inefficient components for a given level of insurance coverage. The traditional view is that moral hazard is wasteful and valued at less than cost. In contrast, we highlight that because insurance effectively offers a way to smoothly finance lumpy expenditures, the socially optimal amount of spending will increase when an individual with liquidity constraints gets access to insurance.

Accounting for liquidity constraints is important when evaluating the value of expanding insurance programs, particularly for low-income populations. Our reevaluation of the value of Medicaid expansion shows that liquidity constrained recipients would receive a much higher benefit than previously estimated. This would affect the normative evaluation of the expansion, underscoring the importance of accurately modeling liquidity constraints in cost-benefit analyses of Medicaid and similar programs.

Incorporating liquidity constraints could be relevant for a range of other empirical analyses and policy recommendations. We show that accounting for liquidity constraints could substantially affect the optimal level of insurance for a population and can change the estimated welfare benefits of vertical insurance choices. In further applications, our results could also be informative about contract design. Accounting for liquidity constraints is necessary to determine the optimal type and level of cost-sharing in many different health insurance programs such that risk-sharing and incentives are balanced optimally. Liquidity constraints may also be related to a literature showing that people respond to the "spot" prices for medical care that they face at a given point in time, rather than merely the effective end-of-year price for that care (e.g., Aron-Dine et al., 2015; Brot-Goldberg et al., 2017). Liquidity constraints could help explain this phenomenon since the spot price rationally matters if people cannot shift money across time. Exploring whether the "spot-price effect" is empirically linked to liquidity constraints and the resulting welfare implications could be a useful direction for future research.

There are also limitations to our approach and extensions that could be the focus of future research. Our hand-to-mouth model is a simple and tractable way of modeling liquidity constraints. We suspect it is a reasonable approximation for many low-income individuals. Yet it also obviously abstracts from the complex realities of both real and perceived access to liquidity. Further work could examine both the role of high borrowing costs as well as behavioral biases, such as present-focus or overconfidence, in shaping spending responses to insurance. Additional empirical evidence

about the nature of access to liquidity and also peoples' beliefs about the availability of liquidity and financing opportunities would also be valuable.

Finally, while prior work on liquidity and moral hazard has focused on unemployment insurance (see Chetty, 2008) and we have extended to health insurance, liquidity constraints are likely important and affect the interpretation of moral hazard in many other insurance domains, such as auto insurance (see Cummins and Tennyson, 1996) and workers' compensation insurance (see Dionne and St-Michel, 1991). Combining data on the liquidity constraints that individuals face with models of insurance in these domains could offer new insights for their optimal design.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263.

Acquatella, A. and Marone, V. (2025). The risk protection value of moral hazard. Working Paper 34156, National Bureau of Economic Research.

Aguiar, M., Bils, M., and Boar, C. (2025). Who are the hand-to-mouth? *Review of Economic Studies*, 92(3):1293–1340.

Aron-Dine, A., Einav, L., Finkelstein, A., and Cullen, M. (2015). Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics*, 97(4):725–741.

Azevedo, E. M. and Gottlieb, D. (2017). Perfect competition in markets with adverse selection. *Econometrica*, 85(1):67–105.

Baily, M. N. (1978). Some aspects of optimal unemployment insurance. *Journal of Public Economics*, 10(3):379–402.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6):2499–2529.

Besley, T. J. (1988). Optimal reimbursement health insurance and the theory of Ramsey taxation. *Journal of Health Economics*, 7(4):321–336.

Brockett, P. L. and Golden, L. L. (1987). A class of utility functions containing all the common utility functions. *Management Science*, 33(8):955–964.

Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318.

Chetty, R. (2006). A general formula for the optimal level of social insurance. *Journal of Public Economics*, 90(10-11):1879–1901.

Chetty, R. (2008). Moral hazard versus liquidity and optimal unemployment insurance. *Journal of Political Economy*, 116(2):173–234.

Chetty, R. and Finkelstein, A. (2013). Social insurance: Connecting theory to data. In Auerbach, A. J., Chetty, R., Feldstein, M., and Saez, E., editors, *Handbook of Public Economics*, volume 5, pages 111–193. Elsevier.

Colarieti, R., Mei, P., and Stantcheva, S. (2024). The how and why of household reactions to income shocks. Technical report, National Bureau of Economic Research.

Cummins, J. D. and Tennyson, S. (1996). Moral hazard in insurance claiming: Evidence from automobile insurance. *Journal of Risk and Uncertainty*, 12:29–50.

De Meza, D. (1983). Health insurance and the demand for medical care. *Journal of Health Economics*, 2(1):47–54.

Dionne, G. and St-Michel, P. (1991). Workers' compensation and moral hazard. *The Review of Economics and Statistics*, 73:236–244.

Dranove, D., Ody, C., and Starc, A. (2021). A dose of managed care: Controlling drug spending in medicaid. *American Economic Journal: Applied Economics*, 13(1):170–197.

Einav, L. and Finkelstein, A. (2018). Moral hazard in health insurance: What we know and how we know it. *Journal of the European Economic Association*, 16(4):957–982.

Einav, L., Finkelstein, A., and Cullen, M. R. (2010). Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877–921.

Einav, L., Finkelstein, A., Ryan, S. P., Schrimpf, P., and Cullen, M. R. (2013). Selection on moral hazard in health insurance. *American Economic Review*, 103(1):178–219.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics*, 5(2):129–151.

Ellis, R. P. and McGuire, T. G. (1990). Optimal payment systems for health services. *Journal of health economics*, 9(4):375–396.

Ericson, K. M. and Laibson, D. (2019). Intertemporal choice. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 1–67. Elsevier.

Ericson, K. M. and Sydnor, J. (2017). The questionable value of having a choice of levels of health insurance coverage. *Journal of Economic Perspectives*, 31(4):51–72.

Ericson, K. M. and Sydnor, J. R. (forthcoming). Liquidity constraints and the value of insurance. *American Economic Journal: Microeconomics*.

Finkelstein, A., Hendren, N., and Luttmer, E. F. (2019). The value of medicaid: Interpreting results from the oregon health insurance experiment. *Journal of Political Economy*, 127(6):2836–2874.

Gross, T., Layton, T. J., and Prinz, D. (2022). The liquidity sensitivity of healthcare consumption: Evidence from social security payments. *American Economic Review: Insights*, 4(2):175–190.

Hendren, N. (2017). Knowledge of future job loss and implications for unemployment insurance. *American Economic Review*, 107(7):1778–1823.

Kaplan, G., Violante, G. L., and Weidner, J. (2014). The wealthy hand-to-mouth. *Brookings Papers on Economic Activity*, 2014(1):77–138.

Layton, T. J., Maestas, N., Prinz, D., and Vabson, B. (2022). Health care rationing in public insurance programs: Evidence from medicaid. *American Economic Journal: Economic Policy*, 14(4):397–431.

Lee, S. C. and Maxted, P. (2023). Credit card borrowing in heterogeneous-agent models: Reconciling theory and data. Technical report, Mimeo.

Marone, V. R. and Sabety, A. (2022). When should there be vertical choice in health insurance markets? *American Economic Review*, 112(1):304–342.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*. Oxford university press New York.

Mukherjee, A., Sacks, D., and Yoo, H. (2024). Does health insurance reduce consumption risk? Evidence from medicaid expansions. Technical report, Mimeo.

Nyman, J. A. (1999a). The economics of moral hazard revisited. *Journal of Health Economics*, 18(6):811–824.

Nyman, J. A. (1999b). The value of health insurance: The access motive. *Journal of Health Economics*, 18(2):141–152.

Nyman, J. A. (2003). *The theory of demand for health insurance*. Stanford University Press.

Nyman, J. A., Koc, C., Dowd, B. E., McCreedy, E., and Trenz, H. M. (2018). Decomposition of moral hazard. *Journal of Health Economics*, 57:168–178.

Olafsson, A. and Pagel, M. (2018). The liquid hand-to-mouth: Evidence from personal finance management software. *The Review of Financial Studies*, 31(11):4398–4446.

Parker, J. A., Souleles, N. S., Johnson, D. S., and McClelland, R. (2013). Consumer spending and the economic stimulus payments of 2008. *American Economic Review*, 103(6):2530–2553.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review*, 58(3):531–537.

Pauly, M. V. and Blavin, F. E. (2008). Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics*, 27(6):1407–1417.

Rowell, D. and Connelly, L. B. (2012). A history of the term "moral hazard". *Journal of Risk and Insurance*, 79(4):1051–1075.

Sergeyev, D., Lian, C., and Gorodnichenko, Y. (2025). The economics of financial stress. *Review of Economic Studies*, 92(5):3403–3437.

Washington, J. (2025). Cleveland Clinic's controversial new policy: pay copays upfront or lose your appointment. *cleveland.com*.

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.

Appendix A Proofs

A.1 Proof of Proposition 1

Proof. Item 1: We must show $m_i^*(\alpha) > m_i^E(\alpha)$ for each type $i \in \{PL, HTM\}$ where $K_{PL} = N$ and $K_{HTM} < N$. We drop the i subscripts on K when it is clear to do so. We abbreviate

$$A(m) = \frac{y - \alpha\pi m}{N} \qquad B_i(m) = \frac{y - \alpha\pi m}{N} - \frac{(1-\alpha)m}{K_i}.$$

We further abbreviate $EU_i(\alpha m_i^E(\alpha))$ as EU_i^E for legibility. To find the efficient level of medical spending we maximize expected utility including the influence of medical spending on the insurance premium. The first order condition for this level is given by

$$\begin{aligned}\frac{\partial EU_i^E}{\partial m} &= \pi h'(m_i^E(\alpha)) - (1-\pi)\alpha\pi u'(A(m_i^E(\alpha))) \\ &\quad - \pi\left[\frac{N-K}{N}\alpha\pi u'(A(m_i^E(\alpha))) + \Big(\frac{\alpha\pi K}{N} + (1-\alpha)\Big)u'(B_i(m_i^E(\alpha)))\right] = 0. \end{aligned} \tag{A1}$$

The second order condition is fulfilled as can be seen by

$$\begin{aligned}\frac{\partial^2 EU_i^E}{\partial m^2} &= \pi\left[(N-K)\Big(\frac{\alpha\pi}{N}\Big)^2 u''(A(m_i^E(\alpha))) + K\Big(\frac{\alpha\pi}{N} + \frac{1-\alpha}{K}\Big)^2 u''(B_i(m_i^E(\alpha)))\right] \\ &\quad + \pi h''(m_i^E(\alpha)) + (1-\pi)\frac{(\alpha\pi)^2}{N}u''(A(m_i^E(\alpha))) < 0. \end{aligned} \tag{A2}$$

We now evaluate $\frac{\partial EU_i^E}{\partial m}$ at $m_i^*(\alpha)$ which is implicitly defined by

$$h'(m_i^*(\alpha)) - (1-\alpha)u'(B_i(m_i^*(\alpha))) = 0 \tag{A3}$$

and obtain

$$\left.\frac{\partial EU_i^E}{\partial m}\right|_{m_i^*(\alpha)} = -\pi^2\alpha\left[\frac{N-K}{N}u'(A(m_i^*(\alpha))) + \frac{K}{N}u'(B_i(m_i^*(\alpha)))\right] - (1-\pi)\alpha\pi\, u'(A(m_i^*(\alpha))) < 0. \tag{A4}$$

Because this expression is negative $EU_i^E(\alpha)$ achieves its maximum to the left of $m_i^*(\alpha)$ and thus $m_i^*(\alpha) > m_i^E(\alpha)$ for all $\alpha \in]0, 1]$.

Item 2: We begin by showing that for $\alpha \in (0, 1)$, $u''' > 0$ is a sufficient condition for $m_i^E(\alpha)$ to be increasing in α. We then shows that this lets us conclude $m_i^E(\alpha) > m_i^*(0)$ and hence $V_i^{MHE}(\alpha) > 0$ for all $\alpha > 0$.

From the implicit function theorem, we can then infer that $sgn\left(\frac{\partial m_i^E(\alpha)}{\partial \alpha}\right) = sgn\left(\frac{\partial^2 EU_i^E}{\partial m \partial \alpha}\right)$. The relevant expression is

$$\begin{aligned}\frac{\partial^2 EU_i^E}{\partial m \partial \alpha} =& -\pi\left[\left(\frac{N-K}{N}\pi + (1-\pi)\right)u'(A(m_i^E(\alpha))) + \left(\frac{K}{N}\pi - 1\right)u'(B(m_i^E(\alpha)))\right] \\ &+ u''(A(m_i^E(\alpha)))m_i^E(\alpha)\pi\left[\pi^2\alpha\frac{N-K}{N^2} + (1-\pi)\pi\alpha\frac{1}{N}\right] \\ &- u''(B(m_i^E(\alpha)))m_i^E(\alpha)\pi\left[\pi\alpha\frac{N-K\pi}{N^2} + (1-\alpha)\frac{N-K\pi}{NK}\right]. \end{aligned} \tag{A5}$$

The first line equals $\pi\left(1-\frac{K}{N}\pi\right)\left(u'(B(m_i^E(\alpha))) - u'(A(m_i^E(\alpha)))\right)$ and is thus positive due to $u'' < 0$. From $u''' > 0$, we know that $u''(B(m_i^E(\alpha))) < u''(A(m_i^E(\alpha)))$. For the second and third line to be positive, it thus remains to be shown that $\pi^2\alpha\frac{N-K}{N^2} + (1-\pi)\pi\alpha\frac{1}{N} \leq \pi\alpha\frac{N-K\pi}{N^2} + (1-\alpha)\frac{N-K\pi}{NK}$. This expression reduces to $0 \leq (1-\alpha)\frac{N-K\pi}{NK}$ which is true for $\alpha \in (0,1)$. Thus $u''' > 0$ is sufficient for $\frac{\partial m_i^E(\alpha)}{\partial \alpha} > 0$.

For full insurance, we can see from (A5) that $\frac{\partial^2 EU_i^E}{\partial m \partial \alpha} = 0$ at $\alpha = 1$. Combined with the previous argument, this implies that $m_i^E(1) \geq m_i^E(\alpha)\ \forall\ \alpha \in [0,1]$.

We know $m_i^E(0) = m_i^*(0)$ since at $\alpha = 0$, the premium term disappears and the efficient objective coincides with the private objective so the efficient FOC collapses to the private FOC. For any $\alpha > 0$, it thus follows that $m_i^E(\alpha) > m_i^E(0) = m_i^*(0)$. Because $m_i^E(\alpha)$ maximizes expected utility for coverage level α, it follows that $EU_i(\alpha, m_i^E(\alpha)) > EU_i(\alpha, m_i^*(0))$ and thus $V_i^{MHE}(\alpha) > 0$.

Item 3: We must show that if Equation (21) holds for type i, then there exists $\hat{\alpha}_i > 0$ such that $V_i^{MH}(\alpha) > 0 \quad \forall \alpha \in (0, \hat{\alpha}_i]$.

We can see from Equation (5) that $V_i^{MH} = 0$ if $\alpha = 0$. Using the definitions of $A(m)$ and $B(m)$ as above and abbreviating $\pi\alpha m_i^*(\alpha)$ as p_α and $\pi\alpha m_i^*(0)$ as p_0, we write

$$\begin{aligned}\frac{\partial V_i^{MH}}{\partial \alpha} =& -\pi\left[\frac{N-K}{N}\frac{\partial p_\alpha}{\partial \alpha}u'(A(m_i^*(\alpha))) + K\left(\frac{1}{N}\frac{\partial p_\alpha}{\partial \alpha} + \frac{1-\alpha}{K}\frac{\partial m_i^*(\alpha)}{\partial \alpha}\right.\right. \\ &\left.\left. -\frac{1}{K}m_i^*(\alpha)\right)u'(B(m_i^*(\alpha)))\right] - (1-\pi)\frac{\partial p_\alpha}{\partial \alpha}u'(A(m_i^*(\alpha))) \\ &+ \pi\frac{\partial m_i^*(\alpha)}{\partial \alpha}h'(m_i^*(\alpha)) + (1-\pi)\pi m_i^*(0)u'(A(m_i^*(0))) \\ &+ \pi m_i^*(0)\left[\frac{N-K}{N}\pi u'(A(m_i^*(0))) + K\left(\frac{\pi}{N} - \frac{1}{K}\right)u'(B(m_i^*(0)))\right] \end{aligned} \tag{A6}$$

We substitute the first-order condition for $m_i^*(\alpha)$ and obtain:

$$\begin{aligned}\frac{\partial V_i^{MH}}{\partial \alpha} =& -\pi\left[\frac{N-K}{N}\frac{\partial p_\alpha}{\partial \alpha}u'(A(m_i^*(\alpha))) + K\left(\frac{1}{N}\frac{\partial p_\alpha}{\partial \alpha} - \frac{1}{K}m_i^*(\alpha)\right)u'(B(m_i^*(\alpha)))\right] \\ &- (1-\pi)\frac{\partial p_\alpha}{\partial \alpha}u'(A(m_i^*(\alpha))) + (1-\pi)\pi m_i^*(0)u'(A(m_i^*(0))) \\ &+ \pi m_i^*(0)\left[\frac{N-K}{N}\pi u'(A(m_i^*(0))) + K\left(\frac{\pi}{N} - \frac{1}{K}\right)u'(B(m_i^*(0)))\right] \end{aligned} \tag{A7}$$

At $\alpha = 0$, $\frac{\partial p_\alpha}{\partial \alpha} = \pi m_i^*(0)$ and thus $\left.\frac{\partial V_i^{MH}}{\partial \alpha}\right|_{\alpha=0} = 0$. Further

$$\begin{aligned}
\frac{\partial^2 V_i^{MH}}{\partial \alpha^2} =& \pi \frac{N-K}{N} \frac{\partial^2 p_\alpha}{\partial \alpha^2} u'(A(m_i^*(\alpha))) + \pi \frac{N-K}{N^2} \left(\frac{\partial p_\alpha}{\partial \alpha}\right)^2 u''(A(m_i^*(\alpha))) \\
& - \pi K \left(\frac{1}{N}\frac{\partial^2 p_\alpha}{\partial \alpha^2} - \frac{1}{K}\frac{\partial m_i^*(\alpha)}{\partial \alpha}\right) u'(B(m_i^*(\alpha))) \\
& + \pi K \left(\frac{1}{N}\frac{\partial p_\alpha}{\partial \alpha} - \frac{m_i^*(\alpha)}{K}\right)\left(\frac{1}{N}\frac{\partial p_\alpha}{\partial \alpha} + (1-\alpha)\frac{1}{K}\frac{\partial m_i^*(\alpha)}{\partial \alpha} - \frac{m_i^*(\alpha)}{K}\right) u''(B(m_i^*(\alpha))) \\
& - (1-\pi)\frac{\partial^2 p_\alpha}{\partial \alpha^2} u'(A(m_i^*(\alpha))) + (1-\pi)\left(\frac{\partial p_\alpha}{\partial \alpha}\right)^2 \frac{1}{N} u''(A(m_i^*(\alpha))) \\
& - \pi \frac{N-K}{N^2} (\pi m_i^*(0))^2 u''(A(m_i^*(0))) - (1-\pi)\frac{1}{N}(\pi m_i^*(0))^2 u''(A(m_i^*(0))) \\
& - \pi K \left(\frac{\pi m_i^*(0)}{N} - \frac{m_i^*(0)}{K}\right)^2 u''(B(m_i^*(0))) \qquad \text{(A8)}
\end{aligned}$$

Observe that $\frac{\partial^2 p_\alpha}{\partial \alpha^2} = 2\frac{\partial m_i^*(\alpha)}{\partial \alpha}\pi + \alpha \frac{\partial^2 m_i^*(\alpha)}{\partial \alpha^2}\pi$. Evaluating at $\alpha = 0$ and again using that in this case $\frac{\partial p_\alpha}{\partial \alpha} = \pi m_i^*(0)$ we can see that

$$\begin{aligned}
\left.\frac{\partial^2 V_i^{MH}}{\partial \alpha^2}\right|_{\alpha=0} =& \pi \frac{\partial m_i^*(\alpha)}{\partial \alpha}\Big[(N - 2K\pi)u'(B(m_i^*(0))) - 2(N - K\pi)u'(A(m_i^*(0))) \\
& - m_i^*(0)\frac{N - \pi K}{K} u''(B(m_i^*(0)))\Big] \qquad \text{(A9)}
\end{aligned}$$

Note that from the implicit definition of $m_i^*(\alpha)$ we can show

$$\frac{\partial m_i^*(\alpha)}{\partial \alpha} = -\frac{u'(B(m_i^*(\alpha))) - (1-\alpha)\frac{m^*(\alpha)}{K} u''(B(m_i^*(\alpha)))}{h''(m_i^*(\alpha)) + \frac{(1-\alpha)^2}{K} u''(B(m_i^*(\alpha)))} > 0. \qquad \text{(A10)}$$

We can thus conclude that $\left.\frac{\partial^2 V_i^{MH}}{\partial \alpha^2}\right|_{\alpha=0} > 0$ if and only if Condition (21) holds. Finally, by continuity of $V_i^{MH,\prime\prime}$, there exists $\hat{\alpha}_i > 0$ such that $V_i^{MH,\prime\prime}(\alpha) > 0$ on $[0, \hat{\alpha}_i]$, hence $V^{MH}(\alpha) > 0$ for all $\alpha \in (0, \hat{\alpha}_i]$.

Item 4: Follows from $V_i = V_i^{RF} + V_i^{MH}$, item 3, and the result that $V_i^{RF} \geq 0$. To see that $V_i^{RF} \geq 0$, note that by virtue of $u'' < 0$, it follows that

$$V_i^{RF}(\alpha) = EU_i(\alpha, m_i^*(0)) - EU_i(0, m_i^*(0)) \tag{A11}$$

$$= \pi \left[(N-K) \left[u\left(\frac{y - \pi\alpha m_i^*(0)}{N} \right) - u\left(\frac{y}{N} \right) \right] \right.$$
$$\left. + Ku\left(\frac{y - \pi\alpha m_i^*(0)}{N} - \frac{(1-\alpha)m_i^*(0)}{K} \right) - u\left(\frac{y}{N} - \frac{m_i^*(0)}{K} \right) \right]$$
$$+ (1-\pi)N \left[u\left(\frac{y - \pi\alpha m_i^*(0)}{N} \right) - u\left(\frac{y}{N} \right) \right] \tag{A12}$$

$$> \alpha m_i^*(0) \left(1 - \frac{\pi K}{N} \right) \left[u'\left(\frac{y}{N} - \frac{m_i^*(0)}{K} \right) - u'\left(\frac{y - \pi\alpha m_i^*(0)}{N} \right) \right] \tag{A13}$$

This expression is positive, because $u'\left(\frac{y}{N} - \frac{m_i^*(0)}{K}\right) > u'\left(\frac{y - \pi\alpha m_i^*(0)}{N}\right)$ by $u'' < 0$.

□

A.2 Proof of Proposition 2

Proof. Item 1: We show that Statement 1, $m_i^*(\alpha) > m_i^E(\alpha)$ for all ($\alpha \in (0, 1]$, continues to hold when $\pi = 1$ for $K \leq N$. Note that in the proof of Proposition 1, the result that $\left.\frac{\partial EU_i^E}{\partial m}\right|_{m_i^*(\alpha)} < 0$ did not rely on $K < N$, but holds for all $K \leq N$. Moreover, $EU_i^E(\alpha, m)$ remains strictly concave in m at $\pi = 1$ (the SOC in the previous proof is unchanged except that the $(1-\pi)$-term drops out, leaving a negative sum since $u'' < 0$ and $h'' < 0$. Therefore the unique maximizer $m_i^E(\alpha)$ must satisfy $m_i^E(\alpha) < m_i^*(\alpha)$. Hence the logic of the previous proof follows through.

Item 2: We show that Statement 2, $u''' > 0$ is a sufficient condition for $m_i^E(\alpha) > m_i^*(0)$ continues to hold when $\pi = 1$ for $K < N$ but does not hold when $K = N$. Examine Equation A5 in the proof of Proposition 1. It shows that $\frac{\partial^2 EU_i^E}{\partial m \partial \alpha} > 0$ when $\pi = 1$ and $K < N$, and hence that $\frac{\partial m_i^E}{\partial \alpha} > 0$. and thus that $m_i^E(\alpha) > m_i^E(0)$ for $\alpha > 0$. By the same logic as Proposition 1, $V_{HTM}^{MHE}(\alpha) > 0$.

However, when $\pi = 1$ and $K = N$, Equation A5 reduces to show that $\frac{\partial^2 EU_i^E}{\partial m \partial \alpha} = 0$ since every term is multiplied by $(N-K)$. This means $m_i^E(\alpha)$ is constant in α. Since $m_{PL}^E(\alpha) = m_{PL}^E(0) = m_{PL}^*(0)$, by definition $V_{PL}^{MHE}(\alpha) = EU(\alpha, m_{PL}^E(\alpha)) - EU(\alpha, m_{PL}^*(0)) = 0$ and there is no efficient moral hazard.

Item 3: We show that Statement 3 still holds for $\pi = 1, K < N$ but that $V_{PL}^{MH}(\alpha) < 0$. As in the proof of Proposition 1 $\left.\frac{\partial V_i^{MH}}{\partial \alpha}\right|_{\alpha=0} = 0$ and $\frac{\partial m_i^*(\alpha)}{\partial \alpha} > 0$ at $\pi = 1$ for $K \leq N$. For the $\pi = 1, K < N$ case, Equation A9 in the previous proof still shows that $\frac{\partial^2 V_{HTM}^{MH}}{\partial \alpha^2}|_{\alpha=0} > 0$ if and only if Condition (21) holds.

However, for the $\pi = 1, K = N$ Condition (21) cannot hold, as it reduces to $-Nu'(B_i)$. As a result, Equation A9 shows that $\frac{\partial^2 V_{PL}^{MH}}{\partial \alpha^2}|_{\alpha=0} < 0$ for $\pi = 1, K = N$ and hence $V_{PL}^{MH}(\alpha) < 0$ for all sufficiently small $\alpha > 0$. To see that $V_{PL}^{MH}(\alpha) < 0$ for all $\alpha \in (0, 1]$, not just locally, note that $V_{PL}^{MH}(\alpha) = EU_{PL}(\alpha, m_{PL}^*(\alpha)) - EU_{PL}(\alpha, m_{PL}^*(0))$. When $\pi = 1$, conditional on chosen m,

EU_{PL} doesn't directly depend on α (since all medical spending will be paid for via premiums and are spread over the same N periods). Since $m^*_{PL}(0) = m^E_{PL}(0)$ maximizes expected utility, and $m *_{PL} (\alpha) \neq m^*_{PL}(0)$, it must be that $EU_{PL}(\alpha, m^*_{PL}(\alpha)) < EU_{PL}(\alpha, m^*_{PL}(0))$ and hence $V^{MH}_{PL}(\alpha) < 0$.

Item 4: $V_i = V_i^{RF} + V_i^{MH}$ For $K < N$, Item 3 shows $V^{MH}_{HTM} > 0$ even when $\pi = 1$. Moreover, Proposition 1 Equation A12 showed that $V^{RF}_{HTM} \geq 0$.

To see that at $\pi = 1, V_{PL}(\alpha) < 0$, note that Item 3 has shown that $V^{MH}_{PL}(\alpha) < 0$. To see that $V^{RF}_{PL} = 0$ at $\pi = 1$, note that Equation A12 for V_i^{RF} from the proof of Proposition 1 reduces to 0 with $K = N, \pi = 1$. □

A.3 Proof of Proposition 3

We consider the first order condition for $m^*_{HTM}(\alpha)$ in Equation (A3) and evaluate it at $m^*_{PL}(\alpha)$. We obtain

$$(1-\alpha)\left[u'\left(\frac{y-\alpha\pi m^*_{PL}(\alpha)-(1-\alpha)m^*_{PL}(\alpha)}{N}\right) - u'\left(\frac{y-\alpha\pi m^*_{PL}(\alpha)}{N} - \frac{(1-\alpha)m^*_{PL}(\alpha)}{K}\right)\right]. \tag{A14}$$

From $\frac{y-\alpha\pi m^*_{PL}(\alpha)-(1-\alpha)m^*_{PL}}{N} > \frac{y-\alpha\pi m^*_{PL}(\alpha)}{N} - \frac{(1-\alpha)m^*_{PL}}{K}$ and $u'' < 0$, we know that the above expression is negative. The second order condition for $m^*_{HTM}(\alpha)$ is fulfilled as can be seen by

$$h''(m^*_{HTM}(\alpha)) + \frac{(1-\alpha)^2}{K}u''\left(\frac{y-\alpha\pi m^*_{HTM}(\alpha)}{N} - \frac{(1-\alpha)m^*_{HTM}}{K}\right) < 0. \tag{A15}$$

Thus we can conclude that $m^*_{PL}(\alpha) > m^*_{HTM}(\alpha)$.

A.4 Proof of Proposition 4

Proof. Item 1: At $\alpha = 1$, it holds that $m^*_{HTM}(1) = m^*_{PL}(1)$, $m^E_{HTM}(1) = m^E_{PL}(1)$ and $EU_{PL}(1, m) = u\left(\frac{y-\pi m}{N}\right) + \pi h(m) = EU_{HTM}(1, m)$. Thus $V^{MH}_{HTM} - V^{MH}_{PL} = V^{MHE}_{HTM} - V^{MHE}_{PL} = EU_{PL}(1, m^*_{PL}(0)) - EU_{PL}(1, m^*_{HTM}(0))$. $EU_{PL}(1, m)$ is concave in m and maximized at $m^E_{PL}(1)$. From $m^E_{PL}(1) > m^*_{PL}(0) > m^*_{HTM}(0)$ it follows that $EU_{PL}(1, m^*_{PL}(0)) - EU_{PL}(1, m^*_{HTM}(0)) > 0$. Note that for $\alpha = 1$ no assumption on u''' is necessary to show that $m^E_{PL}(1) < m^*_{PL}(0)$.

Item 2: From $m^E_{HTM}(1) = m^E_{PL}(1)$ and $m^*_{HTM}(1) = m^*_{PL}(1)$, we can write

$$s^E_{HTM} - s^E_{PL} = \frac{(m^*_{PL}(1) - m^E_{PL}(1))(m^*_{PL}(0) - m^*_{HTM}(0))}{(m^*_{HTM}(1) - m^*_{HTM}(0))(m^*_{PL}(1) - m^*_{PL}(0))} > 0. \tag{A16}$$

The inequality follows from $m^*_{PL}(1) > m^E_{PL}(1)$ (from Proposition 1, item 1), $m^*_{PL}(0) > m^*_{HTM}(0)$ (from Proposition 3), $m^*_{PL}(1) > m^*_{PL}(0)$ and $m^*_{HTM}(1) > m^*_{HTM}(0)$ (both from Equation (A3)). □

A.5 Proof of Proposition 5

Proof. For brevity, we write $A(m) = \frac{y}{N} - \frac{\pi\alpha m}{N}$ and $B(m) = \frac{y}{N} - \frac{\pi\alpha m}{N} - \frac{(1-\alpha)m}{K}$. Further, for any value of $K \leq N$, we write the chosen level of medical care at coverage level α as $m_\alpha(K)$. The chosen level at no insurance coverage is written as $m_0(K)$.

Fix K'. From the statement of the proposition, we know there exists an $\hat{\alpha}(K')$ such that $V^{MH}(\hat{\alpha}(K'), K') = 0$. We further know from that at this point V^{MH} as a function of α crosses the x-axis from above.

Define

$$F(m, K) \equiv EU(\alpha, m; K). \tag{A17}$$

Then

$$V^{MH}(\alpha, K) = F(m_\alpha(K), K) - F(m_0(K), K). \tag{A18}$$

Given the assumption on the health benefit function, $m_\alpha(K)$ is implicitly defined by

$$u'\left(\frac{y}{N} - \frac{\pi\alpha m_\alpha(K)}{N} - \frac{(1-\alpha)m_\alpha(K)}{K}\right) = \frac{\varphi}{1-\alpha} \tag{A19}$$

and $m_0(K)$ is implicitly defined by $u'\left(\frac{y}{N} - \frac{m_0(K)}{K}\right) = \varphi$.

Throughout this argument we evaluate at $(\alpha, K) = (\hat{\alpha}(K'), K')$ and suppress this dependence. Further, for this proof only, we write

$$b := \frac{\pi\alpha}{N}, \qquad d := \frac{1-\alpha}{K'}, \qquad \gamma := b + d, \tag{A20}$$

so that for any m,

$$A(m) = \frac{y}{N} - b\,m, \qquad B(m) = \frac{y}{N} - \gamma\,m, \qquad A(m) - B(m) = d\,m. \tag{A21}$$

Moreover, by the implicit definitions of m_0 and m_α, we define

$$\tilde{m}_0 = \left.\frac{dm_0}{dK}\right|_{K'} = \frac{m_0}{K'}, \qquad \tilde{m}_\alpha = \left.\frac{dm_\alpha}{dK}\right|_{K'} = \frac{(1-\alpha)m_\alpha}{K'^2\,\gamma}. \tag{A22}$$

Lastly, we write partial derivatives of F and V^{MH} as subindices for legibility.

Outline of argument. We begin by showing that V_K^{MH} can be written as an integral over $m \in [m_0, m_\alpha]$. We then sign the integrand. Lastly, we use the sign of V_K^{MH} to show that $\frac{\partial\hat{\alpha}(K')}{\partial K} < 0$.

Step 1: Rewrite V_K^{MH} as an integral over $m \in [m_0, m_\alpha]$. By the chain rule,

$$V_K^{MH} = \Big[F_K(m_\alpha, K') - F_K(m_0, K')\Big] + \Big[F_m(m_\alpha, K')\,\tilde{m}_\alpha - F_m(m_0, K')\,\tilde{m}_0\Big]. \tag{A23}$$

We express each bracket as an integral over the interval $[m_0, m_\alpha]$.

First,

$$F_K(m_\alpha, K') - F_K(m_0, K') = \int_{m_0}^{m_\alpha} F_{mK}(s, K')\, ds. \tag{A24}$$

Second, use the defining property of $\hat{\alpha}(K')$:

$$0 = V^{MH}(\alpha, K') = F(m_\alpha, K') - F(m_0, K') = \int_{m_0}^{m_\alpha} F_m(s, K')\, ds. \tag{A25}$$

Define the linear weight

$$w(s) := \tilde{m}_0 + \frac{s - m_0}{m_\alpha - m_0}\,(\tilde{m}_\alpha - \tilde{m}_0), \tag{A26}$$

so $w(m_0) = \tilde{m}_0$ and $w(m_\alpha) = \tilde{m}_\alpha$. Integrating by parts yields

$$\begin{aligned}
\int_{m_0}^{m_\alpha} F_{mm}(s, K')\, w(s)\, ds &= \Big[F_m(s, K')\, w(s)\Big]_{m_0}^{m_\alpha} - \int_{m_0}^{m_\alpha} F_m(s, K')\, w'(s)\, ds \\
&= F_m(m_\alpha, K')\tilde{m}_\alpha - F_m(m_0, K')\tilde{m}_0 - w' \int_{m_0}^{m_\alpha} F_m(s, K')\, ds \\
&= F_m(m_\alpha, K')\tilde{m}_\alpha - F_m(m_0, K')\tilde{m}_0,
\end{aligned} \tag{A27}$$

where we used that $w'(s)$ is constant and $\int_{m_0}^{m_\alpha} F_m(s, K')ds = 0$. Substituting (A24) and (A27) into (A23) renders

$$V_K^{MH} = \int_{m_0}^{m_\alpha} \Big(F_{mK}(s, K') + w(s)\, F_{mm}(s, K')\Big)\, ds. \tag{A28}$$

Step 2: Sign the integrand. From the definition of expected utility,

$$F(m, K) = (N - \pi K)u(A(m)) + \pi K u(B(m)) + \pi\varphi m, \tag{A29}$$

so, using $\frac{\partial A(m)}{\partial m} = -b$ and $\frac{\partial B(m)}{\partial m} = -\gamma$,

$$F_m(s, K) = -b(N - \pi K)u'(A(s)) - \pi K\gamma u'(B(s)) + \pi\varphi. \tag{A30}$$

This lets us compute

$$F_{mm}(s, K') = (N - \pi K')\, u''(A(s))\, b^2 + \pi K'\, u''(B(s))\, \gamma^2 \tag{A31}$$

and

$$F_{mK}(s, K') = \pi b\big(u'(A(s)) - u'(B(s))\big) \;-\; \pi\,\gamma\,\big(A(s) - B(s)\big)\, u''(B(s)). \tag{A32}$$

Since $A(s) - B(s) = d\,s$, define the linear function

$$r(s) := \frac{A(s) - B(s)}{K'\gamma} = \frac{d\,s}{K'\gamma} = \frac{(1-\alpha)s}{K'^2\gamma}. \tag{A33}$$

Then (A32) can be rewritten as

$$F_{mK}(s,K') = \pi b\big(u'(A(s)) - u'(B(s))\big) \;-\; \pi K'\gamma^2\, u''(B(s))\, r(s). \tag{A34}$$

Now add $w(s)F_{mm}(s,K')$ using (A31) to write the integrand as

$$\begin{aligned} F_{mK}(s,K') + w(s)F_{mm}(s,K') = \pi b\big(u'(A(s)) - u'(B(s))\big) + w(s)(N-\pi K')b^2u''(A(s)) \\ + \pi K'\gamma^2 u''(B(s))\big(w(s) - r(s)\big). \end{aligned} \tag{A35}$$

The first term is strictly negative because $A(s) > B(s)$ and u' is decreasing, hence $u'(A(s)) - u'(B(s)) < 0$. The second term is strictly negative because $w(s) > 0$ and $u''(A(s)) < 0$. For the third term, note that $u''(B(s)) < 0$, so it is non-positive provided $w(s) \geq r(s)$.

Both $w(s)$ and $r(s)$ are linear in s, so for $w(s) \geq r(s)\ \forall\ s \in [m_0, m_\alpha]$ it is sufficient to check the endpoints of the interval $[m_0, m_\alpha]$.

At $s = m_\alpha$, by the derivative above,

$$w(m_\alpha) = m'_\alpha = \frac{(1-\alpha)m_\alpha}{K'^2\gamma} = r(m_\alpha). \tag{A36}$$

At $s = m_0$,

$$w(m_0) = m'_0 = \frac{m_0}{K'} \qquad \text{and} \qquad r(m_0) = \frac{(1-\alpha)m_0}{K'^2\gamma}, \tag{A37}$$

so $w(m_0) \geq r(m_0)$ is equivalent to $K'\gamma \geq 1-\alpha$, which holds because

$$K'\gamma = K'\Big(\frac{\pi\alpha}{N} + \frac{1-\alpha}{K'}\Big) = (1-\alpha) + \frac{\pi\alpha K'}{N} \geq 1-\alpha. \tag{A38}$$

Therefore $w(s) \geq r(s)$ for all $s \in [m_0, m_\alpha]$.

Step 3: Conclude the proof. With $w(s) \geq r(s)$, the third term in (A35) is ≤ 0, while the first two terms are < 0 for all $s \in (m_0, m_\alpha)$. Hence the integrand in (A28) is strictly negative on (m_0, m_α), and thus

$$V_K^{MH}(\hat\alpha(K'), K') = \int_{m_0}^{m_\alpha} \Big(F_{mK}(s,K') + w(s)\, F_{mm}(s,K')\Big)\, ds \;<\; 0. \tag{A39}$$

Now we consider $V^{MH}(\hat\alpha(K'), K') = 0$ and apply the implicit function theorem. It renders:

$$\frac{\partial\hat\alpha(K')}{\partial K'} = -\frac{V_K^{MH}(\hat\alpha(K'), K')}{V_\alpha^{MH}(\hat\alpha(K'), K')}. \tag{A40}$$

We know that $V_K^{MH}(\hat\alpha(K'), K') < 0$. Further, because V^{MH} as a function of α crosses the x-axis from above at K', we know that $V_\alpha^{MH}(\hat\alpha(K'), K') < 0$. This establishes $\frac{\partial\hat\alpha(K')}{\partial K'} < 0$. Because $K_{PL} = N > K_{HTM}$, this establishes the proposition. □

A.6 Proof of Proposition 6

We abbreviate $\frac{y-\alpha\pi m}{N}$ as $A_{PL}(m)$ or $A_{HTM}(m)$, $\frac{y-\alpha\pi m-(1-\alpha)m}{N} = B_{PL}(m)$, and $\frac{y-\alpha\pi m}{N} - \frac{(1-\alpha)m}{K} = B_{HTM}(m)$. Because $h(m)$ is approximately linear, we approximate it as φm. $m_i^*(\alpha)$ is implicitly defined by $\varphi - (1-\alpha)u'(B_i(m_i^*(\alpha))) = 0$. This implies $B_{PL}(m_{PL}^*(\alpha)) = B_{HTM}(m_{HTM}^*(\alpha))$ and thus $m_{PL}^*(\alpha) = \frac{\alpha\pi+(1-\alpha)\frac{N}{K}}{\alpha\pi+(1-\alpha)} m_{HTM}^*(\alpha) < \frac{N}{K} m_{HTM}^*(\alpha)$ which obviously implies $m_{PL}^*(\alpha) > m_{HTM}^*(\alpha)$ for $K < N$.

We abbreviate the premium as $p_i = \alpha\pi m_i^*(\alpha)$ and state the first order condition for the coverage level of individuals with perfect liquidity as

$$\begin{aligned}\frac{\partial EU_{PL}}{\partial\alpha} =& -\pi\frac{\partial p_{PL}}{\partial\alpha}u'(B_{PL}(m_{PL}^*(\alpha))) + \pi m_{PL}^*(\alpha)u'(B_{PL}(m_{PL}^*(\alpha))) - \pi(1-\alpha)\frac{\partial m_{PL}^*(\alpha)}{\partial\alpha}u'(B_{PL}(m_{PL}^*(\alpha)))\\ &+ \pi\frac{\partial m_{PL}^*(\alpha)}{\partial\alpha}\varphi - (1-\pi)\frac{\partial p_{PL}}{\partial\alpha}u'(A_{PL}(m_{PL}^*(\alpha))).\end{aligned} \tag{A41}$$

Substituting the implicit definition of $m_{PL}^*(\alpha)$ and rearranging renders

$$\pi u'(B_{PL}(m_{PL}^*(\alpha))) = \frac{\partial p_{PL}}{\partial\alpha}\frac{\pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A_{PL}(m_{PL}^*(\alpha)))}{m_{PL}^*(\alpha)}. \tag{A42}$$

Similarly, we consider the first order condition for hand-to-mouth individuals and substitute the implicit definition of $m_{HTM}^*(\alpha)$ to obtain

$$\begin{aligned}\frac{\partial EU_{HTM}}{\partial\alpha} =& m_{HTM}^*(\alpha)\pi u'(B_{HTM}(m_{HTM}^*(\alpha))) - \frac{\partial p_{HTM}}{\partial\alpha}\Bigg[\pi\frac{N-K}{N}u'(A_{HTM}(m_{HTM}^*(\alpha)))\\ &+ \pi\frac{K}{N}u'(B_{HTM}(m_{HTM}^*(\alpha))) + (1-\pi)u'(A_{HTM}(m_{HTM}^*(\alpha)))\Bigg]\end{aligned} \tag{A43}$$

We now evaluate $\frac{\partial EU_{HTM}}{\partial\alpha}$ at α_{PL}^*. From $B_{PL}(m_{PL}^*(\alpha)) = B_{HTM}(m_{HTM}^*(\alpha))$:

$$\begin{aligned}\frac{\partial EU_{HTM}}{\partial\alpha}\Bigg|_{\alpha=\alpha_{PL}^*} =& m_{HTM}^*(\alpha)\frac{\partial p_{PL}}{\partial\alpha}\frac{\pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A_{PL}(m_{PL}^*(\alpha)))}{m_{PL}^*(\alpha)}\\ &- \frac{\partial p_{HTM}}{\partial\alpha}\Bigg[\pi\frac{N-K}{N}u'(A_{HTM}(m_{HTM}^*(\alpha))) + \pi\frac{K}{N}u'(B_{HTM}(m_{HTM}^*(\alpha)))\\ &+ (1-\pi)u'(A_{HTM}(m_{HTM}^*(\alpha)))\Bigg]\end{aligned} \tag{A44}$$

The sign of this expression is equivalent to the sign of

$$\begin{aligned}F =& \frac{\frac{\partial p_{PL}}{\partial\alpha}}{m_{PL}^*(\alpha)}\pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A_{PL}(m_{PL}^*(\alpha)))\\ &- \frac{\frac{\partial p_{HTM}}{\partial\alpha}}{m_{HTM}^*(\alpha)}\Bigg[\pi\frac{N-K}{N}u'(A_{HTM}(m_{HTM}^*(\alpha))) + \pi\frac{K}{N}u'(B_{HTM}(m_{HTM}^*(\alpha)))\\ &+ (1-\pi)u'(A_{HTM}(m_{HTM}^*(\alpha)))\Bigg]\end{aligned} \tag{A45}$$

We now note that $\frac{\partial p_i}{\partial \alpha} = \pi m_i^*(\alpha) + \alpha\pi \frac{\partial m_i^*(\alpha)}{\partial \alpha}$. Applying the implicit function theorem to the definitions of the optimal medical spending (noting that second order conditions are fulfilled) renders

$$\frac{\partial m_{PL}^*(\alpha)}{\partial \alpha} = \frac{Nu'(B_{PL}(m_{PL}^*(\alpha))) - (1-\alpha)m_{PL}^*(\alpha)u''(B_{PL}(m_{PL}^*(\alpha)))}{-(1-\alpha)^2 u''(B_{PL}(m_{PL}^*(\alpha)))} \tag{A46}$$

and

$$\frac{\partial m_{HTM}^*(\alpha)}{\partial \alpha} = \frac{Ku'(B_{HTM}(m_{HTM}^*(\alpha))) - (1-\alpha)m_{HTM}^*(\alpha)u''(B_{PL}(m_{HTM}^*(\alpha)))}{-(1-\alpha)^2 u''(B_{HTM}(m_{HTM}^*(\alpha)))}. \tag{A47}$$

We thus know that

$$\frac{\frac{\partial p_{PL}}{\partial \alpha}}{m_{PL}^*(\alpha)} = \pi + \alpha\pi \frac{\frac{N}{m_{PL}^*(\alpha)}u'(B_{PL}(m_{PL}^*(\alpha))) - (1-\alpha)m_{PL}^*(\alpha)u''(B_{PL}(m_{PL}^*(\alpha)))}{-(1-\alpha)^2 u''(B_{PL}(m_{PL}^*(\alpha)))}. \tag{A48}$$

Applying $m_{PL}^*(\alpha) < \frac{N}{K} m_{HTM}^*(\alpha)$ renders

$$\begin{aligned}
\frac{\frac{\partial p_{PL}}{\partial \alpha}}{m_{PL}^*(\alpha)} &> \pi + \alpha\pi \frac{\frac{N}{\frac{N}{K}m_{HTM}^*(\alpha)}u'(B_{PL}(m_{PL}^*(\alpha))) - (1-\alpha)m_{PL}^*(\alpha)u''(B_{PL}(m_{PL}^*(\alpha)))}{-(1-\alpha)^2 u''(B_{PL}(m_{PL}^*(\alpha)))} && \text{(A49)} \\
&= \pi + \alpha\pi \frac{\frac{K}{m_{HTM}^*(\alpha)}u'(B_{PL}(m_{PL}^*(\alpha))) - (1-\alpha)m_{PL}^*(\alpha)u''(B_{PL}(m_{PL}^*(\alpha)))}{-(1-\alpha)^2 u''(B_{PL}(m_{PL}^*(\alpha)))} && \text{(A50)} \\
&= \frac{\frac{\partial p_{HTM}}{\partial \alpha}}{m_{HTM}^*(\alpha)} && \text{(A51)}
\end{aligned}$$

We can apply this relationship to Equationa (A45) and see that

$$\begin{aligned}
F > & \frac{\frac{\partial p_{HTM}}{\partial \alpha}}{m_{HTM}^*(\alpha)}\pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A_{PL}(m_{PL}^*(\alpha))) \\
& - \frac{\frac{\partial p_{HTM}}{\partial \alpha}}{m_{HTM}^*(\alpha)}\left[\pi\frac{N-K}{N}u'(A_{HTM}(m_{HTM}^*(\alpha))) + \pi\frac{K}{N}u'(B_{HTM}(m_{HTM}^*(\alpha)))\right. \\
& \left. + (1-\pi)u'(A_{HTM}(m_{HTM}^*(\alpha)))\right] && \text{(A52)}
\end{aligned}$$

The sign of the right-hand-side is equivalent to

$$\begin{aligned}
G = & \pi u'(B_{PL}(m_{PL}^*(\alpha))) + (1-\pi)u'(A_{PL}(m_{PL}^*(\alpha))) - \left[\pi\frac{N-K}{N}u'(A_{HTM}(m_{HTM}^*(\alpha)))\right. \\
& \left. + \pi\frac{K}{N}u'(B_{HTM}(m_{HTM}^*(\alpha))) + (1-\pi)u'(A_{HTM}(m_{HTM}^*(\alpha)))\right] && \text{(A53)} \\
= & (1-\pi)[u'(A_{PL}(m_{PL}^*(\alpha))) - u'(A_{HTM}(m_{HTM}^*(\alpha)))] \\
& + \pi\frac{N-K}{N}[u'(B_{PL}(m_{PL}^*(\alpha))) - u'(A_{HTM}(m_{HTM}^*(\alpha)))] && \text{(A54)}
\end{aligned}$$

From $m^*_{PL}(\alpha) > m^*_{HTM}(\alpha)$ and $u'' < 0$ it follows that $A_{PL}(m^*_{PL}(\alpha)) < A_{HTM}(m^*_{HTM}(\alpha))$ and thus $u'(A_{PL}(m^*_{PL}(\alpha))) > u'(A_{HTM}(m^*_{HTM}(\alpha)))$. Because $B_{PL}(m^*_{PL}(\alpha)) < A_{PL}(m^*_{PL}(\alpha))$, then also $B_{PL}(m^*_{PL}(\alpha)) < A_{HTM}(m^*_{HTM}(\alpha))$ and thus $u'(B_{PL}(m^*_{PL}(\alpha))) > u'(A_{HTM}(m^*_{HTM}(\alpha)))$. Thus $G > 0$ and in consequence $\frac{\frac{\partial p_{PL}}{\partial \alpha}}{m^*_{PL}(\alpha)} > 0$, implying $\alpha^*_{HTM} > \alpha^*_{PL}$ if we assume an interior solution and thus second-order conditions to hold.

A.7 Proof of Proposition 7

From the first order condition for medical spending in Equation 14, we have that $(1-\alpha)u'(c(\alpha, K)) = h'(m)$ for any α. From the quadratic form of $h(m)$, this implies that: $(1-\alpha)u'(c(\alpha, K)) = a_1 m^*(\alpha) + a_2$.

This gives two linear equations in (a_1, a_2), which can be solved as follows:

$$a_1 = \frac{(1-\alpha_2)u'(c(\alpha_2, K)) - (1-\alpha_1)u'(c(\alpha_1, K))}{m^*(\alpha_2) - m^*(\alpha_1)}, \tag{A55}$$

$$a_2 = (1-\alpha_1)u'(c(\alpha_1, K)) - \frac{(1-\alpha_2)u'(c(\alpha_2, K)) - (1-\alpha_1)u'(c(\alpha_1, K))}{m^*(\alpha_2) - m^*(\alpha_1)} m^*(\alpha_1). \tag{A56}$$

Further, efficient medical spending at α_2 is defined as $m^E_i(\alpha_2) \equiv \arg\max_m \ EU_i(m, p(m, \alpha_2), \alpha_2)$. That is,

$$\begin{aligned} m^E(\alpha_2) = \arg\max_m \pi \left[(N-K)u\left(\frac{y - \pi\alpha_2 m}{N}\right) + Ku\left(\frac{y-\pi\alpha_2 m}{N} - \frac{(1-\alpha_2)m}{K}\right) + h(m)\right] \\ + (1-\pi)Nu\left(\frac{y-\pi\alpha_2 m}{N}\right). \end{aligned} \tag{A57}$$

Taking the FOC and grouping terms, we get:

$$\alpha\frac{N-\pi K}{N}u'\left(\frac{y-\pi\alpha m^E_K(\alpha)}{N}\right) + \left(\frac{\alpha\pi K}{N} + (1-\alpha)\right)u'\left(\frac{y-\pi\alpha m^E_K(\alpha)}{N} - \frac{(1-\alpha)m^E_K(\alpha)}{K}\right) = h'(m). \tag{A58}$$

The left-hand side of the equation is $E[u'_E(\alpha_2, K)]$, given its definition. By the quadratic form then, $E[u'_E(\alpha_2, K)] = a_1 m^E_K(\alpha_2) + a_2$. The proposition follows by substituting and rearranging, giving:

$$\frac{E[u'_E(\alpha_2, K)] - (1-\alpha_1)u'(c(\alpha_1, K))}{(1-\alpha_2)u'(c(\alpha_2, K)) - (1-\alpha_1)u'(c(\alpha_1, K))} = \frac{m^E_K(\alpha_2) - m^*(\alpha_1)}{m^*(\alpha_2) - m^*(\alpha_1)} \tag{A59}$$

A.8 Proof of Proposition 8

We consider both terms in Proposition 8 separately.

From Equation (4) we can write

$$V_i^{RF}(\alpha) = (\pi(N-K)+(1-\pi)N)u\left(\frac{y-\pi\alpha m^*(0)}{N}\right) + \pi K u\left(\frac{y-\pi\alpha m^*(0)}{N} - \frac{(1-\alpha)m^*(0)}{K}\right) - (\pi(N-K)+(1-\pi)N)u\left(\frac{y}{N}\right) - \pi K u\left(\frac{y}{N} - \frac{m^*(0)}{K}\right). \tag{A60}$$

Applying the second-order approximation of u around $\frac{y}{N}$, simplifying and dividing by $u'(y/N)$ renders

$$\frac{V_i^{RF}(\alpha)}{u'(y/N)} = (\pi(N-K)+(1-\pi)N)\left[-\frac{\pi\alpha m^*(0)}{N} - \frac{1}{2}\left(\frac{\pi\alpha m^*(0)}{N}\right)^2 r\right] + \pi K\left[\frac{m^*(0)}{K} + \frac{1}{2}\left(\frac{m^*(0)}{K}\right)^2 r\right] - \pi K\left[\frac{\pi\alpha m^*(0)}{N} + \frac{(1-\alpha)m^*(0)}{K} + \frac{1}{2}\left(\frac{\pi\alpha m^*(0)}{N} + \frac{(1-\alpha)m^*(0)}{K}\right)^2 r\right]. \tag{A61}$$

$$= \frac{1}{2}\pi(2-\alpha)\alpha(m^*(0))^2\left(\frac{1}{K} - \frac{\pi}{N}\right)r. \tag{A62}$$

Which shows the first part of Proposition 8

From Equation (5), we can write

$$V_i^{MH}(\alpha) = (\pi(N-K)+(1-\pi)N)u\left(\frac{y-\pi\alpha m^*(\alpha)}{N}\right) + \pi K u\left(\frac{y-\pi\alpha m^*(\alpha)}{N} - \frac{(1-\alpha)m^*(\alpha)}{K}\right) - (\pi(N-K)+(1-\pi)N)u\left(\frac{y-\pi\alpha m^*(0)}{N}\right) - \pi K u\left(\frac{y-\pi\alpha m^*(0)}{N} - \frac{(1-\alpha)m^*(0)}{K}\right) + \pi(h(m^*(\alpha)) - h(m^*(0))). \tag{A63}$$

Applying the second-order approximation of u around $\frac{y}{N}$, simplifying and dividing by $u'(y/N)$ renders

$$\frac{V_i^{MH}(\alpha)}{u'(y/N)} = (\pi(N-K)+(1-\pi)N)\left[\frac{\pi\alpha m^*(0)}{N} + \frac{1}{2}\left(\frac{\pi\alpha m^*(0)}{N}\right)^2 r\right] + \pi K\left[\frac{\pi\alpha m^*(0)}{N} + \frac{(1-\alpha)m^*(0)}{K} + \frac{1}{2}\left(\frac{\pi\alpha m^*(0)}{N} + \frac{(1-\alpha)m^*(0)}{K}\right)^2 r\right] - (\pi(N-K)+(1-\pi)N)\left[\frac{\pi\alpha m^*(\alpha)}{N} + \frac{1}{2}\left(\frac{\pi\alpha m^*(\alpha)}{N}\right)^2 r\right] - \pi K\left[\frac{\pi\alpha m^*(\alpha)}{N} + \frac{(1-\alpha)m^*(\alpha)}{K} + \frac{1}{2}\left(\frac{\pi\alpha m^*(\alpha)}{N} + \frac{(1-\alpha)m^*(\alpha)}{K}\right)^2 r\right] + \pi(H(m^*(\alpha)) - H(m^*(0))). \tag{A64}$$

$$= \frac{1}{2}\pi(m^*(\alpha) - m^*(0))\left(H'(m^*(\alpha)) + H'(m^*(0)) - 2 - \left[\frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}\right]r(m^*(0) + m^*(\alpha))\right) \tag{A65}$$

Here $H(m) := \frac{h(m)}{u'(y/N)}$ is defined analogously to $H'(m)$ in Appendix C. The key to obtaining the expression is realizing that under the quadratic form, $H(m^*(\alpha)) - H(m^*(0)) = \frac{1}{2}(m^*(\alpha) - m^*(0))(H'(m^*(\alpha)) + H'(m^*(0)))$. We now add $m^E(\alpha) - m^E(\alpha)$ to the first parentheses and rearrange to obatin

$$\begin{aligned}\frac{V_i^{MH}(\alpha)}{u'(y/N)} = &\frac{1}{2}\pi(m^E(\alpha) - m^*(0))\left(H'(m^*(0)) - 1 - \left[\frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}\right] r m^*(0)\right) \\ &- \frac{1}{2}\pi(m^*(\alpha) - m^E(\alpha))\left(1 + \left[\frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}\right] r m^*(\alpha) - H'(m^*(\alpha))\right) \\ &+ \frac{1}{2}\pi(m^*(\alpha) - m^E(\alpha))\left(H'(m^*(0)) - 1 - \left[\frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}\right] r m^*(0)\right) \\ &+ \frac{1}{2}\pi(m^E(\alpha) - m^*(0))\left(H'(m^*(\alpha)) - 1 - \left[\frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}\right] r m^*(\alpha)\right) \quad \text{(A66)}\end{aligned}$$

We now define the last two terms of Equation (A66) as S, $T = \frac{(2-\alpha)\alpha\pi}{N} + \frac{(1-\alpha)^2}{K}$, and $A(m) = H'(m) - 1 - Trm$. From the quadratic form of $h(m)$, we can write $H'(m) = a + bm$. Thus $A(m) = (a-1) + (b - Tr)m$. We can write

$$\begin{aligned}S &= \frac{1}{2}\pi\left[m^*(\alpha)A(m^*(0)) - m^*(0)A(m^*(\alpha)) + m^E(\alpha)\left(A(m^*(\alpha)) - A(m^*(0))\right)\right] &\text{(A67)}\\ &= \frac{1}{2}\pi\left[(a-1)(m^*(\alpha) - m^*(0)) + m^E(\alpha)(b - Tr)(m^*(\alpha) - m^*(0))\right] &\text{(A68)}\\ &= \frac{1}{2}\pi(m^*(\alpha) - m^*(0))\left[(a-1) + m^E(\alpha)(b - Tr)\right] &\text{(A69)}\\ &= \frac{1}{2}\pi(m^*(\alpha) - m^*(0))A(m^E(\alpha)) &\text{(A70)}\end{aligned}$$

Equation (C13) shows that under the given assumptions, the first-order condition for $m^E(\alpha)$ is given as $H'(m^E(\alpha)) = 1 + \left((2-\alpha)\alpha\pi\frac{K}{N} + (1-\alpha)^2\right) r\frac{m^E(\alpha)}{K} = 1 + Trm^E(\alpha)$. Thus $A(m^E(\alpha)) = 0$ and consequently $S = 0$.

Substituting Equations (C10) and (C15) for $H'(m^*(0))$ and $H'(m^*(\alpha))$ in Equation (A66) and simplifying renders

$$\frac{V_i^{MH}(\alpha)}{u'(y/N)} = \frac{1}{2}\pi\left((m^E(\alpha) - m^*(0))\left[\frac{1}{K} - \frac{\pi}{N}\right](2-\alpha)\alpha r m^*(0) - (m^*(\alpha) - m^E(\alpha))\alpha\left(1 + \frac{\pi}{N} r m^*(\alpha)\right)\right). \quad \text{(A71)}$$

We can then substitute s^E to give:

$$\frac{V_i^{MH}(\alpha)}{u'(y/N)} = \frac{1}{2}\pi\alpha(m^*(\alpha) - m^*(0))\left[s^E(0,\alpha)\left(\left(\frac{1}{K} - \frac{\pi}{N}\right)(2-\alpha) r\, m^*(0)\right) - (1 - s^E(0,\alpha))\left(1 + \frac{\pi}{N} r\, m^*(\alpha)\right)\right] \quad \text{(A72)}$$

We can then substitute for $\frac{V_i^{RF}(\alpha)}{u'(y/N)}$ to give the second part of Proposition 8.

Appendix B Comparison of Our Efficient Spending Definition to the Lump-Sum Approach of Nyman (1999a)

Here we compare the approach of Nyman (1999a) to our approach of defining efficient spending, focusing on individuals with perfect liquidity. Nyman (1999a) defines efficient moral hazard as the difference between $m^L(\alpha)$ and $m^*(0)$. For Nyman, m^L is the amount of medical spending optimally chosen by an individual who faced the full marginal cost of spending but who received a lump-sum payment L. The lump-sum payment is designed to capture the income transfer a sick individual receives, and is equal to the amount of medical spending that an insured individual would have chosen, minus premiums paid: $L = (1-\pi)\alpha m^*(\alpha)$. In terms of our model, we can define m^L implicitly through

$$h'(m^L) - u'\left(\frac{y + (1-\pi)\alpha m^*(\alpha) - m^L}{N}\right) = 0. \tag{B1}$$

This concept is not exactly equal to our definition of efficient medical spending, which we term m^E. We exemplify this for full insurance ($\alpha = 1$) and show that the lump-sum approach's spending is higher than the efficient level ($m^L > m^E$). For $\alpha = 1$, m^E is defined by

$$h'(m^E) - u'\left(\frac{y - \pi m^E}{N}\right) = 0. \tag{B2}$$

To see that $m^L > m^E$, evaluate the first-order condition for m^E at m^L. This gives

$$u'\left(\frac{y + (1-\pi)m^*(1) - m^L}{N}\right) - u'\left(\frac{y - \pi m^L}{N}\right). \tag{B3}$$

given that the second-order condition for m^E is fulfilled, $m^L > m^E$ is equivalent to the above expression being negative. Rearranging and realizing that $u'' < 0$ renders

$$\frac{y + (1-\pi)m^*(1) - m^L}{N} > \frac{y - \pi m^L}{N} \tag{B4}$$

This is equivalent to $m^*(1) > m^L(1)$. We know this is true. $m^*(\alpha)$ is defined by $h'(m^*(\alpha)) = 0$ and because $u'\left(\frac{y+(1-\pi)\alpha m^*(\alpha)-m^L}{N}\right) > 0$, it will always be the case that $h'(m^L) > 0$ and thus $m^*(\alpha) > m^L$.

Thus, $m^L > m^E$ which means that the Nyman (1999a) definition overstates the amount of efficient moral hazard compared to our definition. This is because the cash transfer of the lump-sum insurance contract is too large. $m^*(\alpha)$ includes the price effect, which needs to be excluded to calculate the correct amount. Our definition arises naturally, because we simply assume an efficient market without any information asymmetries.

Appendix C Simplified First-Order Conditions for Intuitive Illustration

The following appendix utilizes Taylor approximations to derive linear versions of the first-order conditions used in the results of the papers. These linear approximations are not used to show any additional results, but simply illustrate the mechanisms explored in the paper using simple and thus intuitive functional forms.

C.1 Perfect Liquidity Individuals

We use the first-order condition in Equation (A3) to show how the three levels of medical spending, $m^*_{PL}(0)$, $m^E_{PL}(\alpha)$ and $m^*_{PL}(\alpha)$, can be approximated using linear equations.

We begin with optimal medical spending in the absence of insurance. From

$$u'\left(\frac{y - m^*_{PL}(0)}{N}\right) = h'(m^*_{PL}(0)) \tag{C1}$$

we use a first-order Taylor approximation of $u'(\cdot)$ around the base income $\frac{y}{N}$ to obtain

$$\Rightarrow u'\left(\frac{y}{N}\right) - u''\left(\frac{y}{N}\right)\frac{m^*_{PL}(0)}{N} \approx h'(m^*_{PL}(0)) \tag{C2}$$

$$\Leftrightarrow 1 + r\frac{m^*_{PL}(0)}{N} \approx H'(m^*_{PL}(0)). \tag{C3}$$

Here, r is the coefficient of absolute risk aversion and $H'(m) = h'(m)/u'\left(\frac{y}{N}\right)$ is a normalized version of the marginal utility from health. Because the left-hand side of (C3) is in monetary terms, $H'(m)$ is approximately in monetary terms, too. This is intuitive, because the marginal utility derived from medical spending is divided by that derived from monetary consumption.

For efficient medical spending of insured individuals, we again start at the first-order condition and use a first-order Taylor approximation of $u'(\cdot)$ around the base income $\frac{y}{N}$ to obtain

$$\begin{aligned}
&((1-\alpha)+\pi\alpha)\,u'\left(\frac{y-(1-\alpha)m^E_{PL}(\alpha)-\pi\alpha m^E_{PL}(\alpha)}{N}\right)\\
&+(1-\pi)\alpha u'\left(\frac{y-\pi\alpha m^E_{PL}(\alpha)}{N}\right) = h'(m^E_{PL}(\alpha)) \qquad \text{(C4)}\\
\Rightarrow\ &((1-\alpha)+\pi\alpha)\left[u'\left(\frac{y}{N}\right) - [(1-\alpha)+\pi\alpha]\,u''\left(\frac{y}{N}\right)\frac{m^E_{PL}(\alpha)}{N}\right]\\
&+(1-\pi)\alpha\left[u'\left(\frac{y}{N}\right) - \pi\alpha u''\left(\frac{y}{N}\right)\frac{m^E_{PL}(\alpha)}{N}\right] \approx h'(m^E_{PL}(\alpha)) \qquad \text{(C5)}\\
\Leftrightarrow\ &1 + (1-(1-\pi)\alpha(2-\alpha))\,r\frac{m^E_{PL}(\alpha)}{N} \approx H'(m^E_{PL}(\alpha)) \qquad \text{(C6)}
\end{aligned}$$

Comparing Equations (C3) and (C6), we can see item 2 of Proposition 1 applied to PL individuals, namely that $m^E_{PL}(\alpha) > m^*_{PL}(0)$ for $\alpha > 0$. This is because $1-(1-\pi)\alpha(2-\alpha) < 1$ and we thus

know that the slope for the efficient spending with insurance is lower than that in the case without insurance.

Lastly, we consider medical spending under moral hazard. We again start at the first order condition, apply the approximation and rearrange.

$$(1-\alpha)u'\left(\frac{y-(1-\alpha)m^*_{PL}-\pi\alpha m^*_{PL}}{N}\right)=h'(m^*_{PL}) \tag{C7}$$

$$\Rightarrow (1-\alpha)+\left(1-(2-\pi)\alpha+(1-\pi)\alpha^2\right)r\frac{m^*_{PL}(\alpha)}{N}\approx H'(m^*_{PL}). \tag{C8}$$

Thus, for the linearized definition of $m^*_{PL}(\alpha)$, both the intersection with the y axis and the slope is smaller than for the line defining efficient medical spending. As a result, $m^*_{PL}(\alpha) > m^E_{PL}(\alpha)\ \forall\ \alpha \in]0,1], \pi \in]0,1]$, which is item 1 of Proposition 1 for PL individuals.

C.2 Hand-to-mouth Individuals

The process here is the same as above. For optimal medical spending, we begin at the first-order condition defined in Equation (A3) and apply the first-order Taylor approximation to $u'(\cdot)$ around the base income $\frac{y}{N}$. Thus

$$u'\left(\frac{y}{N}-\frac{m^*_{HTM}(0)}{K}\right)=h'(m^*_{HTM}(0)) \tag{C9}$$

$$\Rightarrow 1+r\frac{m^*_{HTM}(0)}{K}\approx H'(m^*_{HTM}(0)). \tag{C10}$$

Because $K < N$, we know that the slope is steeper for HTM individuals than for PL individuals. This lets us conclude that $m^*_{HTM}(0) < m^*_{PL}(0)$, as is the consequence of Proposition 3.

For the efficient medical spending of insured individuals, we follow the same process as above, using the first-order condition in Equation (A1). Abbreviating $m^E_{HTM}(\alpha)$ as m in the first two lines for legibility, we obtain

$$\begin{aligned}&\frac{N-K}{N}\alpha\pi u'\left(\frac{y-\alpha\pi m}{N}\right)+\left(\frac{\alpha\pi K}{N}+(1-\alpha)\right)u'\left(\frac{y-\alpha\pi m}{N}-\frac{(1-\alpha)m}{K}\right)\\&+(1-\pi)\alpha u'\left(\frac{y-\alpha\pi m}{N}\right)=h'(m)\end{aligned} \tag{C11}$$

$$\begin{aligned}\Rightarrow\ &\frac{N-K}{N}\alpha\pi u'\left(\frac{y}{N}\right)-\frac{N-K}{N}(\alpha\pi)^2u''\left(\frac{y}{N}\right)\frac{m}{N}+\left(\frac{K}{N}\alpha\pi+(1-\alpha)\right)u'\left(\frac{y}{N}\right)\\&-\left(\frac{K}{N}\alpha\pi+(1-\alpha)\right)\alpha\pi u''\left(\frac{y}{N}\right)\frac{m}{N}-\left(\frac{K}{N}\alpha\pi+(1-\alpha)\right)(1-\alpha)u''\left(\frac{y}{N}\right)\frac{m}{K}\\&(1-\pi)\alpha u'\left(\frac{y}{N}\right)-(1-\pi)\alpha^2\pi u''\left(\frac{y}{N}\right)\frac{m}{N}\approx h'(m)\end{aligned} \tag{C12}$$

$$\Leftrightarrow 1+\left((2-\alpha)\alpha\pi\frac{K}{N}+(1-\alpha)^2\right)r\frac{m^E_{HTM}(\alpha)}{K}\approx H'(m^E_{HTM}(\alpha)). \tag{C13}$$

With $K < N$, it is obvious that $(2-\alpha)\alpha\pi\frac{K}{N} + (1-\alpha)^2 < 1$ and thus $m^E_{HTM}(\alpha) > m^*_{HTM}(0)$, as is described in item 2 of Proposition 1.

Lastly, we again consider medical spending under moral hazard. Using the same procedure as above renders

$$(1-\alpha)u'\left(\frac{y-\pi\alpha m^*_{HTM}(\alpha)}{N} - \frac{(1-\alpha)m^*_{HTM}(\alpha)}{K}\right) = h'(m^*_{HTM}(\alpha)) \tag{C14}$$

$$\Rightarrow (1-\alpha) + \left((1-\alpha)\pi\alpha\frac{K}{N} + (1-\alpha)^2\right) r\frac{m^*_{HTM}(\alpha)}{K} \approx H'(m^*_{HTM}(\alpha)) \tag{C15}$$

Interpretations are the same as for the perfect liquidity case. The intersection with the y axis is lower and the slope is smaller because $(2-\alpha)\alpha\pi\frac{K}{N} + (1-\alpha)^2 - \left((1-\alpha)\pi\alpha\frac{K}{N} + (1-\alpha)^2\right) = \alpha\pi\frac{K}{N} > 0$ for $\alpha > 0$. Thus $m^*_{HTM}(\alpha) > m^E_{HTM}(\alpha)\ \forall\ \alpha \in]0,1]$ as is stated in item 1 of Proposition 1.

C.3 Specific scenarios

The two scenarios for certain losses (that is, $\pi = 1$) and full insurance (that is, $\alpha = 1$) are particularly helpful to build intuition. We cover both of them below.

C.3.1 Certain losses: $\pi = 1$

The three medical spending levels for perfect liquidity individuals are described by

$$1 + r\frac{m^*_{PL}(0)}{N} \approx H'(m^*_{PL}(0)) \tag{C16}$$

$$1 + r\frac{m^E_{PL}(\alpha)}{N} \approx H'(m^E_{PL}(\alpha)) \tag{C17}$$

$$(1-\alpha) + (1-\alpha)\, r\frac{m^*_{PL}(\alpha)}{N} \approx H'(m^*_{PL}(\alpha)). \tag{C18}$$

We can easily see that $m^*_{PL}(0) = m^E_{PL}(\alpha)\ \forall\ \alpha$. Further, slope and intersect are lower for medical spending under moral hazard such that $m^*_{PL}(\alpha) > m^E_{PL}(\alpha)$. This implies that for perfect liquidity individuals, under certain losses there is no value of efficient moral hazard, $V^{MHE}_{PL}(\pi = 1) = 0$, while the value of inefficient moral hazard is negative, $V^{MHI}_{PL}(\pi = 1) < 0$ and thus the total moral hazard value will always be negative: $V^{MH}_{PL}(\pi = 1) < 0$.

Considering HTM individuals, we see that

$$1 + r\frac{m^*_{HTM}(0)}{K} \approx H'(m^*_{HTM}(0)) \tag{C19}$$

$$1 + \left((2-\alpha)\alpha\frac{K}{N} + (1-\alpha)^2\right) r\frac{m^E_{HTM}(\alpha)}{K} \approx H'(m^E_{HTM}(\alpha))) \tag{C20}$$

$$(1-\alpha) + \left((1-\alpha)\alpha\frac{K}{N} + (1-\alpha)^2\right) r\frac{m^*_{HTM}(\alpha)}{K} \approx H'(m^*_{HTM}(\alpha)). \tag{C21}$$

This lets us see relatively easily that $m^*_{HTM}(\alpha) > m^E_{HTM}(\alpha) > m_{HTM}(0)$. The latter relationship contrasts the HTM and PL cases.[32] From it, it is obvious that even in the presence of certain losses, there is a positive value of efficient moral hazard for hand-to-mouth individuals, $V^{MHE}_{HTM}(\pi = 1) > 0$. Combined with the negative value of inefficient moral hazard, the sign of the total moral hazard value is undetermined for the hand-to-mouth case.

C.3.2 Full insurance: $\alpha = 1$

The case of full insurance is particularly informative to compare the medical spending of both types of individuals. We restate the definition of medical spending without any insurance as

$$1 + r\frac{m_{PL}(0)}{N} \approx H'(m_{PL}(0)) \tag{C22}$$

$$1 + r\frac{m_{HTM}(0)}{K} \approx H'(m_{HTM}(0)). \tag{C23}$$

Because of $K < N$, we know that $m_{PL}(0) > m_{HTM}(0)$.

Under moral hazard, both types will show the same spending with full insurance

$$0 = H'(m^*_{PL}(\alpha)) = H'(m^*_{HTM}(\alpha)). \tag{C24}$$

This result obtains, because full insurance makes hand-to-mouth individuals act as if they had perfect liquidity. Thus, both types also have the same efficient medical spending as can be seen by

$$1 + \pi r\frac{m^E_{PL}(\alpha)}{N} \approx H'(m^E_{PL}(\alpha)) \tag{C25}$$

$$1 + \pi r\frac{m^E_{HTM}(\alpha)}{N} \approx H'(m^E_{HTM}(\alpha)). \tag{C26}$$

For full insurance, the two types only differ in their counterfactual of no insurance. This difference has implications for the difference in the values of efficient moral hazard. Because all other levels of medical spending are the same, we know that $V^{MHE}_{HTM}(\alpha = 1) > V^{MHE}_{PL}(\alpha = 1)$ and thus $V^{MH}_{HTM}(\alpha = 1) > V^{MH}_{PL}(\alpha = 1)$. This illustrates Proposition 4.

[32] We can see this from $(2 - \alpha)\alpha\frac{K}{N} + (1 - \alpha)^2 < (2 - \alpha)\alpha + (1 - \alpha)^2$.

Appendix D Coverage Elasticity of Demand for Medical Care: Technical Details

Derivation of Equations (25) **and** (26)**.** We consider the general case such that PL results when setting $K_i = N$. We can express Equation (1) as a function of premium and coverage level such that

$$EU_i(p(\alpha), \alpha) = \pi \left[K_i u\left(B_i\right) + h(m(\alpha))\right] + (N - \pi K_i) u\left(A_i\right). \tag{D1}$$

In this case, the optimal insurance coverage is given by $\frac{\partial EU_i}{\partial p}\frac{\partial p}{\partial \alpha} + \frac{\partial EU_i}{\partial \alpha} = 0$. By the envelope theorem, we have $\frac{\partial EU_i}{\partial p} = -\pi \frac{K_i}{N} u'(B_i) - (1 - \pi \frac{K_i}{N}) u'(A_i)$ and $\frac{\partial EU_i}{\partial \alpha} = \pi m_i^*(\alpha) u'(B_i)$. Lastly, the definition of the premium renders $\frac{\partial p}{\partial \alpha} = \pi \left[m_i^*(\alpha) + \alpha \frac{\partial m_i^*(\alpha)}{\partial \alpha}\right]$. We thus have

$$m_i^*(\alpha) u'(B_i) - \left[m_i^*(\alpha) + \alpha \frac{\partial m_i^*(\alpha)}{\partial \alpha}\right] \left[\pi \frac{K_i}{N} u'(B_i) + (1 - \pi \frac{K_i}{N}) u'(A_i)\right] = 0. \tag{D2}$$

Dividing by m_i^* and rearranging renders (26). Setting $K_i = N$ renders (25).

Comparison of markups. For any $\alpha \in (0, 1)$, the right-hand side of equations (25) and (26) can be written as

$$g(a_i, b_i, w_i) = \frac{a_i}{w_i a_i + (1 - w_i) b_i} \tag{D3}$$

with $a_i, b_i, w_i > 0$. Because $A_i > B_i$ and concavity of u, we know that $a_i > b_i$ for all i. By simple derivation, we can then see that

$$\frac{\partial g}{\partial a_i} = \frac{(1 - w_i) b_i}{(w_i a_i + (1 - w_i) b_i)^2} > 0, \tag{D4}$$

$$\frac{\partial g}{\partial b_i} = -\frac{(1 - w_i) a_i}{(w_i a_i + (1 - w_i) b_i)^2} < 0, \tag{D5}$$

$$\frac{\partial g}{\partial w_i} = -\frac{(b_i - a_i) a_i}{(w_i a_i + (1 - w_i) b_i)^2} < 0. \tag{D6}$$

Further $b_{HTM} = u'(A_{HTM}) < u'(A_{PL}) = b_{PL}$ due to Proposition 3. $w_{HTM} = \frac{K_{HTM}}{N}\pi < \pi = w_{PL}$ from $K_{HTM} < N$. Lastly, consider the first order condition for m_i^*: $(1-\alpha)u'(B_i) = h'(m_i^*)$. Because $a_i = u'(B_i)$, $h'' < 0$ and $m_{HTM}^* < m_{PL}^*$, it follows that $u'(B_{HTM}) > u'(B_{PL})$ or equivalently $a_{HTM} > a_{PL}$. All three changes from PL to HTM increase the function g such that the right-hand-side of (26) has to be larger than that of (25) for any positive level of insurance coverage less than 1.

Appendix E Value of Medicaid: Additional Results and Analyses

E.1 Main Analysis: Table of Results

Table E1 – Estimates for the value of Medicaid analysis (main specification)

Variable	$K=12$	$K=11$	$K=10$	$K=9$	$K=8$	$K=7$	$K=6$	$K=5$	$K=4$	$K=3$	$K=2$
Willingness to pay and its decomposition (Figure 6)											
γ	1021	1036	1053	1075	1104	1143	1197	1280	1422	1712	2585
γ^T	566	566	566	566	566	566	566	566	566	566	566
γ^F	0	5	11	18	28	41	59	87	135	235	545
γ^R	13	14	16	18	20	23	27	32	40	53	78
γ^{MH}	442	450	460	473	490	513	545	595	681	858	1396
Moral hazard response and ratio of value to cost (Figure 7, panel (a))											
$\pi(m^*(1)-m^*(0))$	911	911	911	911	911	911	911	911	911	911	911
$\frac{\gamma^{MH}}{\pi(m^*(1)-m^*(0))}$	48%	49%	51%	52%	54%	56%	60%	65%	75%	94%	153%
Efficient level of medical care and efficient share of moral hazard response (Figure 7, panel (b))											
$m^E(1)$	3448	3467	3489	3516	3550	3593	3649	3724	3833	3998	4264
s^E	4%	5%	7%	10%	13%	16%	21%	28%	37%	52%	75%

Note: Source: Authors' calculations from FHL data. Values in the top panel and first rows of the bottom two panels are in \$US. Top panel values correspond to the values for γ_i and its components defined in Equations (27) through (30). Parameters are set as in Table 1 and follow the approach of FHL. We assume that consumption utility has a CRRA utility function with a coefficient of relative risk aversion equal 3. Health production is quadratic with the parameters calibrated from first-order conditions. N is always set to 12. $m^E(1)$ is defined implicitly through $h'(m^E(1)) - u'((y + 0.79\pi m^*(0) - \pi m^E(1))/N) = 0$. s^E is defined as $\frac{m^E(1)-m^*(0)}{m^*(1)-m^*(0)}$.

E.2 Charity Care at Marginal Prices

In their analysis of the value of Medicaid, FHL assume that when uninsured people make their choice of medical care utilization, they know that charity care will cover a fixed share of the cost. This implies that they will only have to pay a portion of the cost, which we denote as $\rho(0)$. Based on the calculations in FHL, it is equal to 0.21. This is in contrast to the share of medical care utilization that people need to pay for under Medicaid, $\rho(1)$, which is 0 because Medicaid provides full coverage.

This assumption does not change the calculation of γ_i or any of its components. Whether people receive a lump-sum transfer of $\psi = 0.79m^*(0)$ or they only pay a share $\rho(0) = 0.21$ of $m^*(0)$ is immaterial to the expected utility of having no insurance as long as $m^*(0)$ is taken from data. As such, Equations (27) through (30) still apply.

What does change, however, is the calibration of the health benefit function $h(m)$. Specifically, the first order condition for determining $m^*(0)$ now takes only the marginal price of medical spending into account. As a result, $a_{2,1} = -a_{1,i}m^*(1)$ still applies, but we now have

$$a_{1,i} = \frac{\rho(0)u'\left(\frac{y}{N} - \frac{\rho(0)m^*(0)}{K_i}\right)}{m^*(0) - m^*(1)}. \tag{E1}$$

The results of the analysis under this charity care assumption are shown in Table E2. Qualitatively, we see the same pattern for γ and its components as K decreases. The overall value increases convexly and this is mostly due to increases in the financing and the moral hazard component of the willingness to pay. Compared to the cost at uninsured prices, γ^{MH} is again at around 50% for individuals with perfect liquidity and over 150% for liquidity constrained individuals with $K = 2$.

Quantitatively, estimates of the value of Medicaid are somewhat lower than those in the main analysis. The reason is simple. The new first order condition leads to less benefit of medical care than the one in equation (31). This lowers the overall value of Medicaid which, after all, is a vehicle of providing medical care. The efficient level of medical care provided in the second to last row of Table E2 shows just how little value people attribute to medical care. The values shown here imply that uninsured individuals with perfect liquidity who got a lump-sum transfer equal to the expected medical care consumption under full coverage would have an efficient medical care consumption lower than $m^*(0)$. This would imply that uninsured individuals overconsume medical care.

Table E2 – Estimates for the value of Medicaid analysis, under charity care at marginal prices

Variable	$K = 12$	$K = 11$	$K = 10$	$K = 9$	$K = 8$	$K = 7$	$K = 6$	$K = 5$	$K = 4$	$K = 3$	$K = 2$
Willingness to pay and its decomposition											
γ	678	686	696	708	724	745	775	821	898	1057	1551
γ^T	566	566	566	566	566	566	566	566	566	566	566
γ^F	0	5	11	18	28	41	59	87	135	235	545
γ^R	13	14	16	18	20	23	27	32	40	53	78
γ^{MH}	99	100	103	106	110	116	123	136	157	204	363
Moral hazard response and ratio of value to the cost at uninsured prices											
$\pi\rho(0)(m^*(1) - m^*(0))$	191	191	191	191	191	191	191	191	191	191	191
$\frac{\gamma_{MH}}{\pi\rho(0)(m^*(1)-m^*(0))}$	51%	53%	54%	55%	58%	60%	65%	71%	82%	106%	189%
Efficient level of medical care and efficient share of moral hazard											
$m^E(1)$	1379	1420	1469	1529	1604	1700	1827	2003	2263	2684	3452
s^E	-176%	-172%	-168%	-163%	-156%	-148%	-137%	-122%	-99%	-63%	4%

Note: Source: Authors' calculations from FHL data. Results in this table assume that uninsured individuals anticipate a fixed share of cost to be covered by charity care such that their marginal price of medical care is $\rho(0) = 0.21$. Values in the top panel and first rows of the bottom two panels are in $US. Top panel values correspond to the values for γ_i and its components defined in Equations (27) through (30). Parameters are set as in Table 1 and follow the approach of FHL. We assume that consumption utility has a CRRA utility function with a coefficient of relative risk aversion equal 3. Health production is quadratic with the parameters calibrated from first-order conditions. N is always set to 12. $m^E(1)$ is defined implicitly through $h'(m^E(1)) - u'((y + 0.79\pi m^*(0) - \pi m^E(1))/N) = 0$. s^E is defined as $\frac{m^E(1)-m^*(0)}{m^*(1)-m^*(0)}$.

E.3 Analyses for $K = 1$

In the main text we refrain from reporting results for $K = 1$. The reason is that FHL impose a consumption floor of $y_{\text{floor}} = \$1977$ which implies a maximum per-period out-of-pocket spending of \$627. For $K = 1$ the out-of-pocket spending of uninsured sick people would be \$715 in the consumption period with a loss and would thus exceed this maximum. To report estimated values for Medicaid for $K = 1$, we would thus have to make adjustments to the model. Below we discuss three possible adjustments and the results they imply.

The first adjustment is to lower y_{floor} such that it allows for an out-of-pocket payment of \$715 in a single consumption period. The adjustment is sizable in that the new value is $y_{\text{floor}} = \$924$, more that \$1000 less than the value in FHL. The results for the value of Medicaid both for $K = 12$ and $K = 1$ are displayed in the first two columns of Table E3. The perfect liquidity case is not affected, because y_{floor} is never reached in the binary model. For $K = 1$, we can see an extremely high value of Medicaid at \$8,604.

Table E3 – Value of Medicaid for $K = 1$ under various adjustments

Variable	Lower y_{floor}		Higher y		Lower $m^*(0)$	
	$K = 12$	$K = 1$	$K = 12$	$K = 1$	$K = 12$	$K = 1$
Willingness to pay and its decomposition						
γ	1021	8604	1022	7911	978	6958
γ^T	566	566	566	566	497	497
γ^F	0	5412	0	3568	0	2836
γ^R	13	139	12	129	10	111
γ^{MH}	442	2486	443	3648	471	3513
Moral hazard response and ratio of value to cost						
$m^*(1) - m^*(0)$	911	911	911	911	980	980
$\frac{\gamma_{MH}}{m^*(1)-m^*(0)}$	48%	273%	49%	400%	48%	358%
Efficient level of medical care and efficient share of moral hazard						
$m^E(1)$	3448	4553	3444	4542	3340	4536
s^E	4%	100%	3%	99%	2%	98%

Note: Source: Authors' calculations from FHL data. Values in the top panel and first rows of the bottom two panels are in \$US. Top panel values correspond to the values for γ_i and its components defined in Equations (27) through (30). Parameters are set as in Table 1 and follow the approach of FHL. We assume that consumption utility has a CRRA utility function with a coefficient of relative risk aversion equal 3. Health production is quadratic with the parameters calibrated from first-order conditions. N is always set to 12. $m^E(1)$ is defined implicitly through $h'(m^E(1)) - u'((y + 0.79\pi m^*(0) - \pi m^E(1))/N) = 0$. s^E is defined as $\frac{m^E(1)-m^*(0)}{m^*(1)-m^*(0)}$. Compared to the main analysis, the first two columns set an annual consumption floor of \$924. The second two columns increase income to $y = 10,556$. The last two columns decrease uninsured medical care consumption to $m^*(0) = \$627$.

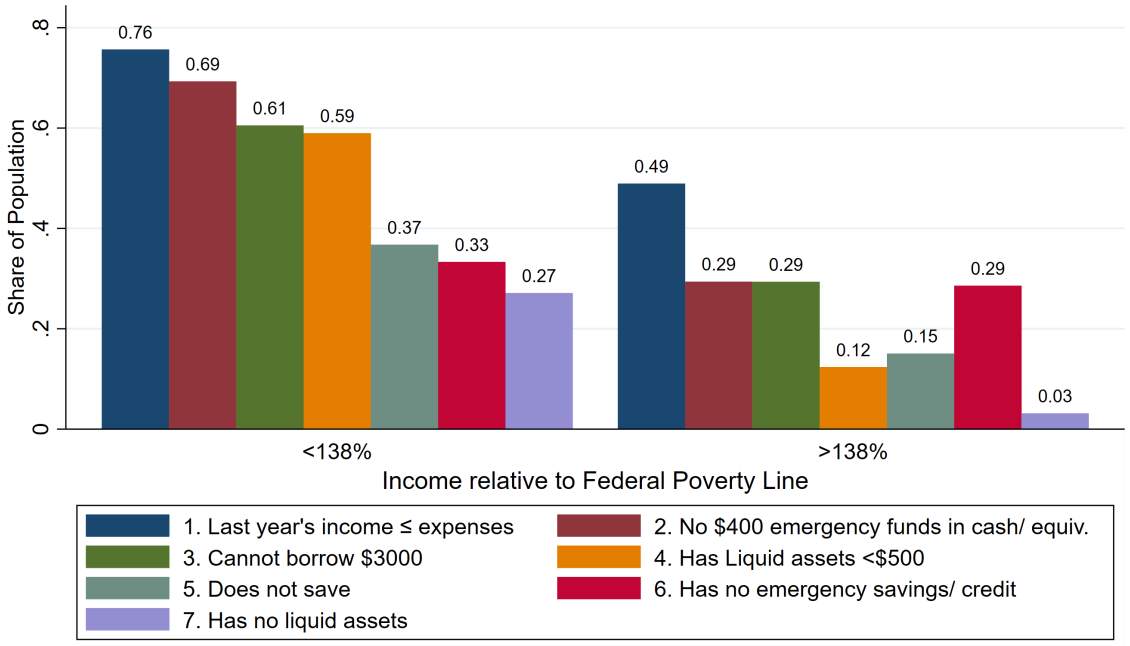
The second adjustment is to set income y so high that the out-of-pocket spending of \$715 in one consumption period reduces consumption exactly to the consumption floor imposed by FHL. This implies $y = \$10,556$. The results are displayed in columns three and four of Table E3. The perfect liquidity case is only affected slightly. Because the income is higher, the calibrated health

benefit function puts a slightly higher emphasis on health over consumption. The effect is miniscule, however. For $K = 1$, we can again see an extremely high value of Medicaid at \$7,911.

Lastly, we can reduce $m^*(0)$ such that it implies an out-of-pocket spending of \$627. This is achieved by setting $m^*(0) = \$3,317$. Notably, we hold the absolute value of the lump-sum transfer at the same level as for the other simulations to make the settings as comparable as possible. The results are displayed in the last two columns of Table E3. The perfect liquidity case is again only affected slightly. Lower $m^*(0)$ reduces the transfer value of insurance 1:1. At the same time, there is now a stronger reaction of medical care consumption to increased insurance coverage. The higher values of $m^*(1) - m^*(0)$ increases the value of moral hazard. The overall effect is a modest reduction in the willingness to pay for Medicaid of about \$43. For $K = 1$ we can again also see an extremely high value of Medicaid at \$7,067. Even though this is the smallest value of all three adjustments, it is nevertheless very high when compared to results for lower values of K.

Appendix F Proxies for Hand-to-Mouth Status

Figure F1 – Hand-to-Mouth Indicators by Income Relative to the Federal Poverty Line



Note: Data are from a pooled analysis of the 2013, 2016, 2019 and 2022 waves of the Survey of Consumer Finances (SCF, N = 85,785) except for the question on having emergency savings or credit, which was only asked in 2022, 2019, and 2016 (N = 53,267). Data on having $400 in emergency funds which is taken from the 2023 wave of the Survey of Household Economics and Decisionmaking (SHED, N = 8,390).

It is difficult to get estimates of the fraction of people who are liquidity constrained, and more explicitly, who act as if they are hand-to-mouth. We therefore get a range of empirical proxies. To do so, we turn to two different surveys. First, we use the Survey of Consumer Finances (SCF), pooling across the 2013, 2016, 2019 and 2022 waves, totaling N=17,157. We supplement this with the data on having $400 in emergency funds which is taken from the 2023 wave of the Survey of Household Economics and Decisionmaking (SHED, N = 8,390). We want to describe the experience of low-income individuals who are most likely to be eligible for Medicaid. We therefore split the sample based on whether the individual is in a household that is above or below 138% of the Federal Poverty Level. Data on the federal poverty line for the respective years is taken from the Poverty Guidelines for 48 Contiguous States provided by the Office of the Assistant Secretary for Planning and Evaluation.[33]

Figure F1 displays the fraction of the sample displaying each of the following criteria:

[33] In the SHED, a household is categorized as being below 138% of the poverty line, if it reports an income in a category where the lower bound is below this threshold. This slightly understates the prevalence of liquidity indicators for poorer households.

1. Has zero or negative savings in the last calendar year (in blue)
2. Would not pay an unexpected $400 expense in cash or equivalent (in maroon),
3. Answered negatively to the question "In an emergency could you (or your {husband/wife/partner}) get financial assistance of $3,000 or more from any friends or relatives who do not live with you?" (in green)
4. Has liquid assets less than $500, defined as the sum of reported (market) values of all checking accounts, saving accounts, certificates of deposit, money market accounts, mutual funds, bonds, and stocks owned by the household (in yellow),
5. Answers "Don't save" to the question "Which of the following best describes your saving habits?" (in gray),
6. Answers "Postpone bills", "get an extra job", or "other" to the question "If tomorrow you experienced a financial emergency that left you unable to pay all of your bills, how would you deal with it? Would you borrow money, would you spend out of savings or investments, would you postpone paying bills, work more or get an extra job, or would you do something else?" (in red),
7. Has no liquid assets, defined as the sum of reported (market) values of all checking accounts, saving accounts, certificates of deposit, money market accounts, mutual funds, bonds, and stocks owned by the household (in lavender)

Appendix G Additional Details for the Selection on Moral Hazard Analysis

Table G1 summarizes the parameters used for the simulation. The initial source is Azevedo and Gottlieb (2017). We simulate individuals that are log normally distributed with the moments provided in the table. r is the coefficient of absolute risk aversion, H is the moral hazard parameter, μ is the mean of the loss distribution, and σ is the standard deviation of the loss distribution.

Table G1 – Parameters Used for Simulating Selection on Moral Hazard Exercise

	r	H	μ	σ
Mean	1.0E-4	1330	4340	7739
Log covariance				
r	0.25	-0.01	-0.12	0
H		0.28	-0.03	0
μ			0.20	0
σ				0.25

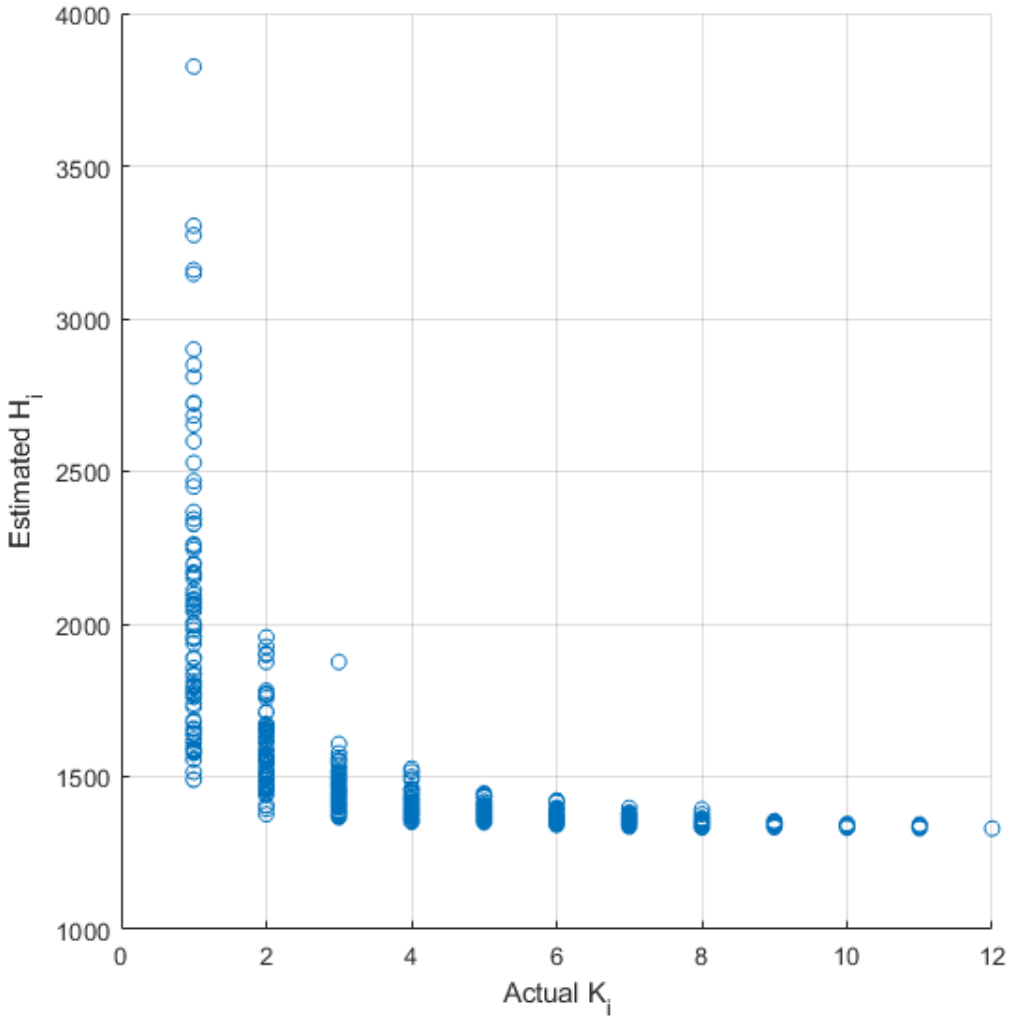
Note: We simulate individuals that are log normally distributed with the moments provided in this table. r is the coefficient of absolute risk aversion, H is the moral hazard parameter, μ is the mean of the loss distribution, and σ is the standard deviation of the loss distribution. After all parameters are drawn, we set H_i to its mean value of 1330 for all individuals.

In comparison to Azevedo and Gottlieb (2017) we make three changes. First and by design, after the parameters were drawn, we set $H_i = 1330$ for all individuals. Second, Azevedo and Gottlieb (2017) use a low coefficient of absolute risk aversion, because the one estimated in Einav et al. (2013) produces unreasonably high willingness to pay for insurance. With such a low risk aversion coefficient, however, a CARA utility function will be almost linear and not react to liquidity constraints at all. We thus increase risk aversion by a factor of 10 which still leaves it an order of magnitude smaller than the estimated value of Einav et al. (2013), but in a comparable range to other estimates of the parameter (e.g., Marone and Sabety, 2022). Third, in order to maintain the willingness to pay for insurance at reasonable levels, we divide the original value of σ by the square root of 10. In the absence of moral hazard, the willingness to pay for insurance is $\mu + \frac{1}{2}r\sigma^2$. This modification thus decreases the willingness to pay back to the reasonable levels Azevedo and Gottlieb (2017) were aiming for.

Figure G1 shows the parameters estimated by the naïve econometrician assuming that $K_i = N \; \forall \; i$. Each dot represents one individual. The estimation only recovers H_i correctly if $K_i = N$. In all other cases, the estimated value is higher than the actual value. This is because with $K_i < N$, the individual is more reactive in their medical spending than would be predicted for perfect liquidity. The estimated model takes this increased reactivity as evidence for higher levels of H_i. For lower values of K_i the heterogeneity in the estimated H_i increases. This is due to differences in r_i which

do play a role in determining $m_i^*(l;\alpha)$ in the data generating model, but not in the model that assumes perfect liquidity.

Figure G1 – Scatter of Estimated H_i Parameters Against Liquidity Constraints K_i



Note: Plots the distribution of estimated H_i parameters using Equation (34), under the assumption of perfect liquidity ($K = N$). When $K = 12$, the model recovers the true parameter H_i. When $K < 12$, estimated H_i is more variable and exceeds the true H_i.