THE WELFARE COST OF UNPRICED HETEROGENEITY IN INSURANCE MARKETS

By Valentino Dardanoni and Paolo Li Donni

We consider the welfare loss of unpriced heterogeneity in insurance markets, which results when private information or regulatory constraints prevent insurance companies to set premiums reflecting expected costs. We propose a methodology which uses survey data to measure this welfare loss. After identifying some 'types' which determine expected risk and insurance demand, we derive the demand and cost functions for each market defined by these unobservable types, quantifying the efficiency costs of unpriced heterogeneity. We apply our methods to the US Long-Term Care and Medigap insurance markets, where we find that unpriced heterogeneity causes substantial inefficiency.

JEL Classification Numbers D82, G22, I11.

Keywords Unpriced Heterogeneity, Welfare Loss, Selection, Multidimensional Private Information, Long-Term Care Insurance, Medigap.

1. Introduction

In many insurance markets private information or regulatory constraints prevent insurers to set premiums reflecting individuals' costs. In this context, *unpriced heterogeneity* refers to all characteristics which affect insurance demand and expected claims, but are not priced by insurance companies. Under unpriced heterogeneity, individuals pay insurance premiums which do not reflect expected costs. Since efficiency requires that individuals should purchase insurance if and only if their willingness to pay for the contract is greater than their expected costs, the existence of unpriced heterogeneity implies that some individuals overbuy and some underbuy. How large is the resulting welfare loss? If contracts were priced conditional on individual expected risks, or if the social planner could use appropriate observables to price heterogeneity, how large

Date: October 2, 2013.

⋆ Valentino Dardanoni: Università di Palermo, Dipartimento SEAF, 90129, Palermo, Italy, valentino.dardanoni@unipa.it. Paolo Li Donni: Università di Palermo, Dipartimento SEAF, Viale delle Scienze, 90129, Palermo, Italy, paolo.lidonni@unipa.it. We thank Alberto Bennardo, Alberto Bisin, Liran Einav, Antonio Forcina, Mark Machina, Andrea Pozzi, Dan Silverman, Joel Sobel, Antonio Tesoriere, seminar audiences at CORE Louvain, Imperial College, LSE, UCSD, University of York, University of Lancatser, CRENoS University of Cagliari, EIEF Rome, Health Econometrics workshop at Coimbra, and the 2011 Petralia Workshop for helpful comments. Nicola Persico deserves special thanks for his time and encouragement.

would the welfare gain be? In this paper, we propose a methodology to address these questions.

To measure welfare loss, we recover willingness-to-pay and expected costs for each participant in the market by segmenting the market into 'types' that have costs independent of coverages. To identify these unobservable types, we exploit as a source of external variation individual observable characteristics which act as indicators of risk and risk preferences, but are not used by insurers' to price the contract (called "unused observables" by Finkelstein and Poterba [2006]). If these types were contractible, they could be segregated into separate 'synthetic' markets, each with its own insurance contract. Within each synthetic market individuals have the same expected claims, and all heterogeneity is idiosyncratic; there would be no welfare loss from charging the same price to all customers. The estimates of willingness-to-pay and expected costs allow us to calculate counterfactual price changes conditional on the features of the unobserved markets, allowing prices to vary with expected costs. The magnitude of the welfare gain that could be achieved pricing these types depends on demand slopes and on the difference between current and counterfactual efficient prices in each market through the usual deadweight loss triangle.

The literature on the empirical appraisal of the welfare implications of pricing insurance contracts is recent (see Einav, Finkelstein and Levin [2010] for a survey). Our paper is mostly related to the papers of Einav, Finkelstein and Cullen [2010] and Bundorf, Levin and Mahoney [2012].[1] The seminal paper of Einav, Finkelstein and Cullen [2010] studies the welfare cost of private information in a simple and intuitive framework where the researcher describes preferences for insurance and expected utilization estimating the aggregate demand and marginal and average cost curves. Bundorf, Levin and Mahoney [2012] consider a structural model of health plan choice taking into account unobservable health risk, which interact in the estimation of the demand

[1]Recent papers closely related with ours are also Lustig [2011] on the efficiency of Medigap insurance and Geruso [2012] on unpriced heterogeneity in health plan choice. These studies analyze how heterogeneous preferences over insurance –uncorrelated with individual insurable risk– can induce, under a uniform price setting, inefficient self-sorting into plans.

and expected cost functions, and measure the welfare gain of allowing prices to vary with expected costs.

Our paper differs from current literature in several aspects. First, compared with Einav, Finkelstein and Cullen [2010], we decompose the aggregate market into separate synthetic markets. Within these markets, we find efficient prices and measure the inefficiency of the implicit cross-subsidization —which do not emerge by looking at the aggregate market— giving a very different perspective on the nature of the distortions created by unpriced heterogeneity. In comparison with Bundorf, Levin and Mahoney [2012], who model private risk as a univariate normally distributed unobserved variable and assume idiosyncratic preference heterogeneity, we allow for nonparametric structural heterogeneity jointly affecting both claims and preferences.

The other difference between the current paper and existing ones is that we show how our approach can also be applied when the researcher uses survey data rather than firm data with exogenous variation in prices. For our purpose, the advantage of using survey data is twofold. First, and this is key to our approach, survey data typically contains information on individual characteristics which act as indicators of riskiness and insurance preferences, helping to extract systematic unobserved differences in claims and coverages. Secondly, survey data allow to explore the efficiency of insurance markets with large representative samples, enabling the study of many cases where the researcher does not have access to appropriate firm or administrative data. On the other hand, the most relevant limitation of our approach is precisely due to the use of survey data, which generally do not contain information on individual insurance premiums. To sidestep this issue, we do a calibration exercise using external information on the price elasticity of the aggregate demand for insurance taken from recent literature. In general, the researcher may leverage a set of credible estimates of the policy effects of interest using reasonable ranges for this parameter.

We apply our methods to the US Long-Term Care and Medigap insurance markets. In both markets we find that unpriced heterogeneity causes substantial inefficiency. In LTC insurance we find that, conditional on insurers' risk classification, there are two underlying unobserved synthetic markets with large differences in insurance valuation and expected risks. High-risk individuals are almost 2.5 times more likely to use a nursing home than low-risk ones: we estimate that a one-dollar subsidy to the high-risk types costs the low-risks between 3.1 and 5.5 dollars. In our sample we also find that the welfare loss of current pricing –compared to pricing at expected costs in each synthetic market– ranges from 7.5 to 10 percent of total coverage cost.

In the Medigap market, where premiums are heavily regulated, estimation suggests the existence of substantial structural heterogeneity which can be cross-classified into types with striking differences in estimated insurance and medical care choices. In our sample, the welfare loss of current pricing –compared to pricing at expected costs in each synthetic market– ranges between 14 and 28 percent of total coverage costs; a one-dollar subsidy to the high risks costs between 1.6 and 2.6 dollars to the low risks. We also find that optimally set uniform (non-differentiated) prices reduces the estimated welfare loss very marginally, suggesting that the loss of allocative efficiency of current pricing is for the most part due to the lack of price differentiation.[2]

The paper is organized as follows: in the next section we describe a simple model of insurance markets under unpriced heterogeneity. Sections 3 and 4 discuss how to measure social efficiency using survey data. In section 5 we discuss identification and estimation of the synthetic markets. Sections 6 and 7 present our two applications. Section 8 concludes. Section 9 collects all tables and figures used in the main text.

2. The insurance market under unpriced heterogeneity

Consider an insurance market where individuals can buy a given insurance contract which protects them from a probabilistic loss. Firms offer the contract at a price p which depends on some observable characteristics $\boldsymbol{x}$, and individuals make a binary

[2]It is worth noting that our modeling approach ignores a potential benefit of unpriced heterogeneity: the cross-subsidy generated by the lack of price discrimination could transfer resources to those with a relatively higher marginal utility of consumption. In fact, our measure of the welfare cost of the cross-subsidization of high risks at the expense of low risk types is upwardly biased if high risk types have a higher marginal utility of consumption. Results of Finkelstein, Luttmer and Notowidigdo [2009] suggest that those with higher medical utilization do not in fact have a higher marginal utility of consumption.

choice of whether to buy the contract.[3] Let I denote a binary variable which takes value 1 if an individual has bought a contract which protects her from a fixed loss with monetary value D, and L a binary variable which takes value 1 if the individual incurs the loss. This is a good setting when using survey data, where typically the researcher only observes whether the individual occurred the loss and whether she is covered by an insurance plan. To ease notation, unless explicitly mentioned we set $D = 1$, so that the loss probability equals expected costs, but we relax this assumption whenever appropriate.

To describe the market, we assume, as in Einav, Finkelstein and Cullen [2010], that consumers make a discrete choice to buy the contract, and risk neutral providers set prices in a Nash equilibrium. To keep notation simple, we initially condition on a given value of the observable characteristics $\boldsymbol{x}$ used to set prices. Thus, all individuals face the same price and are undistinguished by insurers.

In this context, residual *unpriced heterogeneity* refers to all variables which affect claims and coverage after conditioning on $\boldsymbol{x}$. Such variables include not only unobserved private information considered in standard insurance models, but possibly any other characteristic (observable or not) which cannot be used either by regulatory laws or industry norms. Thus, individuals in the market may differ in several dimensions with respect to both risk factors and risk preferences.

We assume that the population is made of a discrete number of 'types' $t \in \{1, \ldots, M\}$. Notice that the label of the types is arbitrary, and no order is assumed on the types.[4] In this context, different types are simply meant to capture significant residual heterogeneity without any assumption on its underlying structure.

Conditional on a given value of the pricing variables $\boldsymbol{x}$, letting T denote the discrete random variable in {1,...,M} which classifies the different types, the observed joint

[3]Compare with most of the theoretical literature which follows the seminal Rothschild and Stiglitz [1976] analysis, which endogenizes not only the pricing of insurance contracts but also the level of coverage.

[4]For example, if types differ by two binary variables capturing risk type and risk preference, as shown by Smart [2000] there is no ordering of the four types since single crossing of the indifference curves fails to hold.

distribution of (I, L) is the aggregation of a finite mixture of M unobserved joint distributions such that, conditional on T, I and L are *independent*:

$$P(I, L) = \sum\nolimits_{t=1}^{M} P(T = t)P(I \mid t)P(L \mid t). \tag{1}$$

In this sense, types extract all systematic variation in cost and insurance demand, and all residual heterogeneity can be considered as idiosyncratic. In each market t realized costs and taste for insurance may differ individually, but expected cost are constant and unrelated to individuals willingness to buy insurance (since L and I are conditionally independent). In other words, within each type, there is no selection.

For each type t there is an unobserved 'synthetic' market, with insurance demand function $q_t(p) = P(I = 1 \mid t, p)$, and constant (expected) average and marginal costs $AC_t = MC_t = P(L = 1 \mid t) = c_t$. The observed aggregate demand for insurance is $q(p) = \sum_t P(T = t)q_t(p)$. The observed aggregate (expected) total cost curve is $TC(p) = \sum_t P(T = t)q_t(p)c_t$, where, as stressed by Einav, Finkelstein and Cullen, aggregate costs are determined by the costs of the sample of individuals who endogenously choose to buy the contract. If identical risk neutral insurance providers set prices in a Nash equilibrium, the equilibrium price p^* is the solution to $q(p) = AC(p) = \frac{\sum_t P(T=t)q_t(p)c_t}{\sum_t P(T=t)q_t(p)}$.

2.1. **A graphical example.** Suppose there are two heterogeneous types which can be segmented into synthetic markets such that within each market the MC curve is flat and equals the AC curve. These two (unobserved) markets are depicted in Figure 1, where the horizontal axis indicates the proportion of individual who buy the contract. Individuals in the first market have on average *both* a higher probability of buying the contract and to experience the loss compared with the individuals of the second market. Assume there is an equal number of individuals in the two markets. Since by assumption individuals in both markets have the same characteristics $\boldsymbol{x}$, they all face the same price, namely the zero-profit equilibrium price denoted p^* in the figure.

Figure 1 about here

If the two groups were contractible, there would be two independent markets with different (flat) MC curves, and price differentiation according to expected costs would generate social efficiency, since social efficiency requires that individuals buy the contract if and only if their willingness to pay exceeds their expected cost. The welfare loss from unpriced heterogeneity (the inefficiency of actual pricing compared to optimal discriminatory prices) is equal to the weighted sum of the deadweight loss triangles CEF in the two figures.

The welfare loss is the result of the implicit cross-subsidization of high risks at the expense of low ones caused by the uniform price. Notice that while by construction the financial gain of cross subsidization for high risk types 1 equals the financial loss for low risk types 2 (the areas GFCB are equal), the welfare gain of types 1 is smaller than the welfare loss of types 2 (compare the areas BCEG in the two figures). Thus, we can also appraise the (in)efficiency of cross-subsidization as the ratio between the welfare loss of low risks and the welfare gain of high risks (the weighted ratio of the areas BCEG in the two figures). This gives a measure of the welfare cost of transferring one dollar from the low risk to the high risk types.

3. Welfare loss measures with survey data

To measure the welfare loss of unpriced heterogeneity of a given insurance contract we first need to decompose the observed aggregate market into the efficient synthetic markets defined by the M types. To this end, we can exploit as a source of external variation individual characteristics which act as observable preference and cost indicators but are not used by insurance companies to price the contracts, which are often available in survey data. Examples of such variables are wealth, cognitive abilities, occupational risk, risk reducing or increasing behavior such as preventive care, seat belt use, smoking and drinking or, if panel data are available, past insurance choices and claims.

Suppose that, conditional on the variables $\boldsymbol{x}$ used by insurers to price contracts, we observe the aggregate probability of buying the insurance contract and filing a claim $P(I)$ and $P(L)$, and want to estimate $P(I \mid T = t)$ and $P(L \mid T = t)$ together with the types probabilities $P(T = t)$, for $t = 1, \ldots, M$. Suppose we have a set of insurance preference and risk indicators $Z_1, \ldots, Z_H$. The key identifying assumption is that $Z_1, \ldots, Z_H$, jointly with I and L, are conditionally independent given types:

$$P(I, L, Z_1, \ldots, Z_H) = \sum_t P(T = t) \cdot P(I \mid t) \cdot P(L \mid t) \cdot P(Z_1 \mid t) \cdots P(Z_H \mid t). \quad (2)$$

This assumption, known as local independence in the finite mixture literature, basically says that if we knew an individual's type, then knowing the value of any of the variables $I, L, Z_1, \ldots, Z_H$ would not help predict the value of any other one. In other words, all residual heterogeneity is idiosyncratic, and the partition by types captures all systematic unpriced heterogeneity.

The local independence assumption (2) implies that if we have a sufficient number of indicators we can estimate the unobserved probabilities $P(I \mid T = t)$, $P(L \mid T = t)$, $P(Z_h \mid T = t)$ and $P(T = t)$ using the observed moments $P(I, L, Z_1, \ldots, Z_H)$. For example, if the indicators Z_h are binary variables, letting $S = H + 2$, we have 2^S observable moments, $S \cdot M$ unobserved conditional probabilities and $M - 1$ types marginal probabilities. Thus equation (2) can be seen as a nonlinear system of $2^S - 1$ equations into $(S + 1) \cdot M - 1$ unknowns, which can be nonparametrically estimated. Necessary and sufficient conditions for identification are discussed in section 5.1.

3.1. **Approximate welfare loss.** Suppose we have an estimate of the proportion of types $P(T = t)$, the demand for insurance $P(I = 1 \mid t)$ and expected costs $P(L = 1 \mid t)$ for all unobserved synthetic markets $t = 1, \ldots, M$. The equilibrium price p^* is equal to average cost (e.g. in a perfectly competitive or contestable market with no administrative costs)

$$p^* = P(L = 1 \mid I = 1) = \frac{\sum_t P(T = t) q_t(p^*) c_t}{\sum_t P(T = t) q_t(p^*)} \quad (3)$$

where $q_t(p^*)$ denotes the estimated proportion of t-type individuals who buy the contract at the equilibrium price and c_t denotes the t-types' estimated expected claims.

The estimated quantities $q_t(p^*)$, c_t, $P(T = t)$ and p^*, coupled with appropriate external information on the price elasticity of the aggregate insurance demand, allow an approximate measure of the welfare cost of unpriced heterogeneity in each synthetic market. In market t, insurance demand at the current equilibrium price p^* is equal to the (estimated) quantity $q_t(p^*)$. Taking a first order expansion of the demand curve $q_t(p)$ around p^* and letting $\Delta_t = c_t - p^*$, we approximate the (unknown) efficient quantity $q_t(c_t)$ as $q_t\big(p^* + \Delta_t\big) \sim q_t(p^*) + \Delta_t q_t'(p^*)$. Letting η_t denote the price elasticity of demand at $\big(p^*, q_t(p^*)\big)$ and rearranging terms, the welfare loss from unpriced heterogeneity in market t is approximately equal to

$$WL_t = -\frac{1}{2}\eta_t \frac{q_t(p^*)}{p^*}\Delta_t^2, \tag{4}$$

which can be compared with the usual formula for the deadweight loss of taxes and subsidies.

Under the assumption that in each market t the demand functions at the current price have approximately the same slope, η_t is equal to $\eta\frac{q(p^*)}{q_t(p^*)}$ and approximate welfare loss is

$$WL = -\frac{1}{2}\eta \frac{q(p^*)}{p^*} \sum\nolimits_t P(T = t)\Delta_t^2, \tag{5}$$

which can be calculated from estimated claims and insurance probabilities in each synthetic market, togheter with external information on the aggregate insurance demand elasticity η.

3.2. **Remarks.**

(i) When D is different from 1, without further information on p^* or D welfare calculations cannot be given a monetary value. However, since varying D both consumer surplus and costs are multiplied by the same factor, we can still calculate welfare loss *relative* to other measures which use the same estimated demand and cost functions. In particular, we will scale welfare loss to total cost, and the social welfare of an inefficient allocation to the socially efficient one.

(ii) When estimating p^* from conditional claims and coverage probabilities (see equation (3)), one needs to assume that p^* equals average cost excluding administrative costs. In the presence of administrative costs (in general the costs of running an insurance company), $p^* = \lambda \cdot \frac{\sum_t P(T=t) q_t(p^*) c_t}{\sum_t P(T=t) q_t^*} \cdot D$ for some $\lambda > 1$. In this case we can assume that the marginal expected costs which are relevant for efficiency calculations are inclusive of running costs, that is, the relevant marginal expected cost of selling an insurance contract to t-types is $EC(T = t) = \lambda \cdot c_t \cdot D$, so that zero-profits are defined to include running costs, but, for the purpose of calculating relative welfare loss nothing of substance is changed.

(iii) These calculations are computed for a given value of the conditioning variables $\boldsymbol{x}$. In practice, one can either calculate these measures at a given value of $\boldsymbol{x}$ (such as the average or the median value), or calculate them for all combinations of $\boldsymbol{x}$ in the sample and average them out.

(iv) We do not use exogenous price variation –that is, price variation independent on observed characteristics $\boldsymbol{x}$– for deriving our welfare loss measure. While this is a key characteristic of our approach, it means that we have to rely on external information of the price elasticity of the demand of insurance, and our welfare calculations depend on external calibration through η. The researcher may experiment with calibration of different values of η taken from different papers in the same market, other literature in similar markets, or simply informed guess, leveraging out different estimates for appraising the robustness of estimated welfare loss.

(v) The insurance products we consider in this paper are purchased in a binary choice. In this setting, researchers often estimate logit or probit demand specifications. To the extent we think these functional forms are a better approximation of demand, we can use them to derive the welfare loss from non-uniform pricing in closed form. In appendix A we discuss a standard utility model such as that employed in Bundorf, Levin and Mahoney [2012], where the (money metric) utility of buying insurance for a type-t individual is

$$u_t = w_t - p + \sigma\epsilon, \tag{6}$$

with ϵ being an i.i.d. extreme value idiosyncratic individual preference shifter and the mean utility of not buying the contract is normalized to zero.[5] In appendix A we show how we can use estimated insurance and claims probabilities across synthetic markets, jointly with external information on aggregate price elasticity η, to derive the parameters w_t and σ, and how to use the estimated structural parameters to calculate consumer surplus for an exact welfare loss estimation.

The logit specification is theoretically sound and has some flexibility since it does not force neither equal slopes nor equal elasticities at the equilibrium price across the synthetic markets. However, while the logit specification is common in this literature, it is still restrictive since variation in the elasticity of demand is not identified by synthetic market-level estimates from the data. For example, with constant σ synthetic markets with the same equilibrium quantity have the same demand elasticity. In section 5 below we discuss how the assumption of constant σ across synthetic markets can be tested using variation in $\boldsymbol{x}$.

3.3. **Conditional independence.** The decomposition of the observed market into synthetic markets with given demand and constant MC curves is key for our policy analysis. The decomposition is based on the assumption that the heterogeneity variable T extracts all systematic differences in individuals' valuation of the insurance contract and expected costs; from this assumption it follows that if the types were contractible and heterogeneity was accordingly priced, welfare loss would disappear.

It is important to realize that for efficiency what counts is that types with the same expected costs face the same price; residual heterogeneity in willingness to pay within types with the same expected cost is not important. For example, if there are two types with the same expected costs but different risk preferences, but we lump them

[5]Compared to Bundorf, Levin and Mahoney [2012] we use a less restrictive utility formulation, since unpriced heterogeneity is flexibly and nonparametrically modeled by a (multidimensional) characterization of types (in contrast, Bundorf, Levin and Mahoney define structural heterogeneity only in terms of risk, with a normal distribution specification).

together into a single type, pricing at the common marginal cost would still eliminate welfare loss.[6] On the other hand, if two types have different costs and demands but we lump them together, it is theoretically possible that pricing the lumped type at their expected costs does not increase efficiency (see e.g. the example in footnote 3 in Einav and Finkelstein [2011]).

This underscores the importance of uncovering all types with structural claims and coverage heterogeneity. A theorem by Suppes and Zanotti [1981] shows that any joint distribution of binary variables can be decomposed into a mixture of conditionally independent tables for a large enough value of M. In practice only a small number of types M is identified in a given model, and thus it is important to test the conditional independence property for the chosen value M. This assumption can be tested by estimating M association parameters of the joint distribution of (I, L) for each t. In particular, we can estimate the *log-odds ratio* between I and L for each type $t = 1, \dots, M$, and test their significance by a LR test statistic which is asymptotically distributed as χ^2_M. In practice, depending on the nature of the sample, the econometric procedure extracts types which give a statistically sufficient fine partition of the unpriced heterogeneity, and it may be useful to check the robustness of the policy implications to the number of synthetic markets M, as we show in our applications.

4. Social welfare revisited

Letting $CS_t(p) = \int_p^\infty q_t(s)ds$ denote consumer surplus for the t-types, social welfare under a given uniform price p is

$$W(p) = \sum\nolimits_t P(T = t)\big(CS_t(p) - (c_t - p)q_t(p)\big). \tag{7}$$

Social welfare in the zero-profit market equilibrium is $W^m = W(p^*)$. The socially efficient allocation is obtained when the planner can segment individuals in each market;

[6]However, as we discuss in section 4 below, neglecting risk preference may bias welfare calculations under nonlinear insurance demand. Geruso [2012] also argues that taking into account systematic differences in willingness to pay is important for the optimal design of insurance contracts.

in each market t the planner maximizes $W_t(p_t) = CS_t(p_t) - (c_t - p_t)q_t(p_t)$, with solution $\bar{p}_t = c_t$, $t = 1, \ldots, M$. The efficient social welfare is $W^e = \sum_t P(T = t)CS_t(c_t)$.

In the efficient allocation the planner knows each individual's type. If the planner knows the population type distribution but does not have any information on the type distribution of the individuals in the population, a *constrained efficient* allocation can be obtained by imposing the uniform price $\bar{p}$ which is the solution to the maximization of (7), say by appropriate taxes or subsidies. Denoting $W^{ce} = W(\bar{p})$, clearly $W^e \geq W^{ce} \geq W^m$.

The key intuition of Einav, Finkelstein and Cullen [2010] (see also Einav and Finkelstein [2011]) is that the constrained efficient allocation –and an estimation of the welfare loss of the market equilibrium compared with the constrained efficient allocation– can be obtained from the *aggregate* functions which describe the market, namely the demand, AC and MC curves. Einav, Finkelstein and Cullen show how these aggregate functions can be straightforwardly estimated using data on insurance purchase, claims and exogenously varying prices (that is, price variation independent on individual characteristics $\boldsymbol{x}$). Furthermore, estimating the aggregate MC curve allows a simple test for detecting selection in the market: if the aggregate MC curve is downward (upward) sloping the market is adversely (favourably) selected, and there is no selection if and only if the aggregate MC curve is flat.[7]

The Einav, Finkelstein and Cullen approach is clear and illuminating. However, aggregate market functions may not uncover the true unobserved heterogeneity and its welfare effects. For example, in the model described in figure 1, at each price there is a constant proportion of high and low risk individuals purchasing the contract, which implies that the aggregate AC curve is flat and identical to the MC curve. Estimating the aggregate demand and cost curves would suggest absence of selection and market

[7]As clearly explained by Einav, Finkelstein and Cullen [2010], the logic of the test is straightforward: under adverse selection, higher risk individuals have higher demand for insurance, so moving down the aggregate demand curve the insurance market is populated by increasingly lower risk individuals, which is reflected in a downward sloping MC curve (the opposite holds under favorable selection).

efficiency (the optimal uniform price is equal to the market price and $W^m = W^{ce}$), even in the presence of large market inefficiency.[8]

While this is a rather special theoretical example, using aggregate market functions may be problematic in real situations. Recent empirical literature (e.g. Cutler, Finkelstein and McGarry [2008]) has emphasized the relevance of the multidimensionality of private information. When unpriced heterogeneity is multidimensional, not only the definition of selection becomes blurred, but also looking at the aggregate market to detect selection can be very misleading. For example, suppose there is systematic heterogeneity along two binary dimensions, namely expected cost and risk tolerance, so that there are four types of individuals in the population. In the appendix B we illustrate the four resulting synthetic markets using logit demands, and the observed aggregate market under the assumption that there is an equal number of individuals in each type. Since aggregate expected costs come from a mixture of high and low risk individuals, the aggregate MC curve is equal to the AC curve even in the presence of substantial market inefficiency.[9] Note that this example is not an implausible theoretical curiosity, but is actually close in spirit to the empirical decomposition that we document below in our analysis of Medigap insurance (see also Dardanoni and Li Donni [2012]).

Under multidimensional heterogeneity, estimation of welfare effects may be unreliable also when the aggregate market is segmented only by risk, neglecting risk preferences. Figure 5 appendix B illustrates the two markets which result after aggregation of the four underlying synthetic markets (fig. 4) by risk types. In this example, estimated welfare loss segmenting by risk only is overestimated by 42% compared to true welfare loss. To understand what is going on here, consider the structural utility equation (6), and suppose we decompose the types by riskness r_t and risk aversion ra_t:

$$u_t = r_t + ra_t - p + \sigma\epsilon$$

[8]This example relies on the use of linear demand with equal intercept in the two markets.
[9]What is happening in this example is that joining the two (adversely selected) markets on the left, the resulting aggregate marginal cost is decreasing; joining the two (favourably selected) markets on the right, the aggregate marginal cost curve is increasing. The overall observable MC curve is flat.

If we estimate this model with logit idiosyncratic preference ϵ neglecting risk attitudes as in Bundorf, Levin and Mahoney [2012], eventhough ra_t is uncorrelated with ϵ, we will end up overestimating the scale parameter σ and thus estimating a less elastic demand function with higher welfare loss. This follows from the nonlinearity of the demand function, and is analogous to the well known problem of uncorrelated neglected heterogeneity in non linear models (see e.g. Wooldridge [2002]).

4.1. **$\boldsymbol{z}$-conditional allocation.** A relevant intermediate case between the efficient (where the planner knows each individual type) and the constrained efficient (where the planner knows only the population type distribution) allocation is when the planner may estimate the type probability distribution for each individual. Let $P(T = t \mid i)$, $t = 1, \dots, M$, $i = 1, \dots, N$ be the type distribution for a population of N individuals. Then the planner can choose a vector of prices $\boldsymbol{p} = [p_1, \dots, p_N]$ to maximize

$$W(\boldsymbol{p}) = \sum\nolimits_i \sum\nolimits_t P(T = t \mid i)\big(CS_t(p_i) - (c_t - p_i)q_t(p_i)\big).$$

The planner can use a relevant subset $\boldsymbol{z}$ of the observable indicators which help predicting individuals' type. If the planner knows how the types probability distribution depends on $\boldsymbol{z}$, can set individual prices depending on $\boldsymbol{z}$. Let $\boldsymbol{z}_i$ denote the value of $\boldsymbol{z}$ for individual i and $P(T = t \mid \boldsymbol{z}_i)$, $t = 1, \dots, M$ his type probability distribution. Social welfare is equal to

$$W(p(\boldsymbol{z})) = \sum\nolimits_i \sum\nolimits_t P(T = t \mid \boldsymbol{z}_i)\big(CS_t(p(\boldsymbol{z}_i)) - (c_t - p(\boldsymbol{z}_i))q_t(p(\boldsymbol{z}_i))\big) \tag{8}$$

where $p(\boldsymbol{z})$ denotes the vector of prices $p(\boldsymbol{z}_1), \dots, p(\boldsymbol{z}_N)$.

Choosing an appropriate price function $\bar{p}(\boldsymbol{z})$ to maximize (8) will result in the $\boldsymbol{z}$-conditional social welfare $W^z = W(\bar{p}(\boldsymbol{z}))$. If $\boldsymbol{z}$ determines uniquely the types (i.e. if each value of $\boldsymbol{z}$ maps into a degenerate type distribution), then social welfare will be equal to the efficient allocation. If the types distribution is independent on $\boldsymbol{z}$, social welfare will be equal to the constrained efficient allocation. In most cases, $\boldsymbol{z}$-conditional social welfare will lie between these two extremes.

5. Identification and estimation of the synthetic markets

5.1. **Identification.** Conditional on $\boldsymbol{x}$, which we omit for simplicity, the key parameters we need for our welfare analysis are

$$P(L=1 \mid T=t),\ P(I=1 \mid T=t),\ P(T=t),\ t=1,\dots,M, \tag{9}$$

so that we want to identify $2 \times M + M - 1$ unobservable parameters. Since we observe only the joint distribution of I and L with 3 free parameters, unless covariates' variation and parametric restrictions are imposed, it is impossible to identify (9) even with $M = 2$.

As discussed above, our strategy is to look for external sources of variation by means of the set of observable indicators Z_h, $h = 1, \dots, H$, which allow to increase the number of observable parameters; and at the same time to impose appropriate conditional independence assumptions which reduce the number of parameters to be identified.[10]

For simplicity, we assume that chosen indicators Z_h are actually binary variables. This restriction aims to simplify the discussion, but our econometric analysis can be performed as long as these indicators are discrete. To simplify notation let us rename the response variables $I, L, Z_1, \dots, Z_H$ as $Y_1, \dots, Y_S$, so that $S = H + 2$. The joint distribution of the observed responses is a function of the unobserved types' marginal probabilities and responses' conditional probabilities:

$$P(Y_1 = i_1, \dots, Y_S = i_S) = \sum\nolimits_{t=1}^{M} P(T=t) \prod\nolimits_{j=1}^{S} P(Y_j = i_j \mid T=t),\ i_j \in \{0,1\}. \tag{10}$$

Equation (10) shows that the moments of the distribution of $I, L, Z_1, \dots, Z_H$ are linked to the unobserved parameters by a nonlinear system of $2^S - 1$ equations into $(S+1) \times M - 1$ unknowns. If a sufficient number of indicators are available, a large number of types can be identified. For example, in the LTCI application $S = 6$, and in principle up to 9 different types can be identified. In the Medigap application $S = 13$, so in principle a very large number of types can be identified.

It is well known that counting the number of observable parameters is only a necessary condition for identification of model parameters; there are pathological examples

[10] The use of conditional independence assumptions as a possible strategy for achieving nonparametric identification is discussed, among others, in the survey by Matzkin [2007].

in the literature on finite mixture models which show that this counting condition is not sufficient. More formally, if we write compactly equation (10) above as $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the vector which arrays the unobservable parameters of the mixture model and $\boldsymbol{q}$ the vector which arrays the observable joint distribution of the responses, model (14-15-16-17) is identified if, for every $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$,

$$\boldsymbol{f}(\boldsymbol{\theta}') = \boldsymbol{f}(\boldsymbol{\theta}'') \rightarrow \boldsymbol{\theta}' = \boldsymbol{\theta}'' \tag{11}$$

that is, the model is identified if the mapping between the parameters of the observed variables and the mixture model parameters is one-to-one; in other words, the model is identified if the equation $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{\theta})$ has at most one solution for each value of $\boldsymbol{q}$.

While the state of the mathematical art is such that there are no general conditions for the uniqueness of the solution of nonlinear systems of equation, Allman, Matias and Rhodes [2009] provide sufficient conditions for the *generic identifiability* of finite mixture models. In particular, in Corollary 5 they show generic identifiability of finite mixtures of binary variables whenever the number of observed variables S and the number of mixing types M satisfies

$$S \geq 2\lceil \log_2(M) \rceil + 1 \tag{12}$$

where $\lceil x \rceil$ denotes the smallest integer at least as large as x. Since generic identifiability implies that the set of nonidentifiable parameters has measure zero, generic identifiability of a model is sufficient for statistical inference, since observed data has probability one of coming from an identifiable distribution. Equation (12) implies that, for all values of $\boldsymbol{x}$, in the LTCI application below at least 4 types are identified, while in the Medigap application at least 64 types are identified.

5.2. **Estimation.** When the variables in $\boldsymbol{x}$ are discrete and sufficient observations are available for each configuration, we can nonparametrically estimate the parameters in (10) in each stratum by simply setting $E(\boldsymbol{q} - \boldsymbol{f}(\boldsymbol{\theta})) = \mathbf{0}$. Using observed sample probabilities, $\boldsymbol{\theta}$ can be found solving for example an appropriate nonlinear least squares problem.

When $\boldsymbol{x}$ is continuous or each strata contains too few observations, to estimate these probabilities for each vector $\boldsymbol{x}$ we assume that $P(I = 1 \mid \boldsymbol{x}, t)$ and $P(L = 1 \mid \boldsymbol{x}, t)$ are linear in $\boldsymbol{x}$ and T:

$$P(I = 1 \mid \boldsymbol{x}, t) = F(\alpha_I(t) + \boldsymbol{x}'\boldsymbol{\beta}_I), \quad P(L = 1 \mid \boldsymbol{x}, t) = F(\alpha_L(t) + \boldsymbol{x}'\boldsymbol{\beta}_L) \tag{13}$$

for some appropriate link function F (we use the logit link).

Before discussing how to estimate the model parameters α's and β's, we pause to comment on the assumptions behind insurance demand in (13). The insurance demand equation in (13) follows from the logit utility model (6) by adding a linear term for $\boldsymbol{x}$. As argued above, the utility model (6) uses the assumption that σ is constant across markets. This assumption (which is common in the literature) is quite restrictive, but is unavoidable since the model is identified off the aggregate parameter η. When we estimate the model conditional on $\boldsymbol{x}$, we can use variation in $\boldsymbol{x}$ to test this assumption. In particular, if $u_t = w_t + \boldsymbol{x}'\boldsymbol{\beta}_I(t) - p + \sigma_t \epsilon$, then $\alpha_I(t) = \frac{w_t - p}{\sigma_t}$ and $\boldsymbol{\beta}_I(t) = \frac{\boldsymbol{\beta}_I}{\sigma_t}$, so that we can estimate insurance demand letting $\boldsymbol{\beta}_I(t)$ vary across markets, and use a standard LR test statistic for the linear restrictions $\boldsymbol{\beta}_I(1) = \cdots = \boldsymbol{\beta}_I(M)$.

5.3. **A discrete multivariate finite mixture model.** To estimate the parameters of interest for each synthetic market t we use a *discrete multivariate finite mixture model* which is composed of four parts:

(1) The demand for insurance equation

$$P(I = 1 \mid \boldsymbol{x}, t) = F(\alpha_I(t) + \boldsymbol{x}'\boldsymbol{\beta}_I). \tag{14}$$

(2) The expected cost equation

$$P(L = 1 \mid \boldsymbol{x}, t) = F(\alpha_L(t) + \boldsymbol{x}'\boldsymbol{\beta}_L). \tag{15}$$

(3) An auxiliary system of equations which is instrumental to achieve identification of the heterogeneous types:[11]

$$\begin{array}{rcl} P(Z_1 = 1 \mid t) & = & F(\alpha_{Z_1}(t)), \\ \vdots & \vdots & \vdots \\ P(Z_H = 1 \mid t) & = & F(\alpha_{Z_H}(t)). \end{array} \tag{16}$$

(4) The types membership probabilities. To force the types probabilities to lie between zero and one and sum to one, it is convenient to use a multinomial logit parameterization:

$$P(T = t) = \frac{\exp(\alpha_T(t))}{\sum_{t=1}^{M} \exp(\alpha_T(t))}, \ \alpha_T(M) = 0. \tag{17}$$

The $M - 1$ logit parameters α_T are simply reparameterizations of the membership probabilities, and do not impose any parametric restriction on the distribution of T.

The discrete multivariate finite mixture model is defined by equations (14-15-16-17), with α's and β's being the model parameters. Model (14-15-16-17) can be seen as an instance of a discrete multivariate MIMIC model (Joreskog and Goldberger [1975]). Contrary to the MIMIC model, the unobserved heterogeneity T is not a continuous univariate variable on the real line, but an unstructured nonparametric variable. The unstructured nature of the random effects $\alpha_I(t)$ and $\alpha_L(t)$ in equations (14)-(15) is well suited to capture the possibly multidimensional residual heterogeneity. Our model can be seen as a multivariate extension of the single equation *semiparametric finite mixture model* (see e.g. Follmann and Lambert [1989] for a binary logit example) which has become popular in economics after the seminal paper by Heckman and Singer [1984].[12]

[11]Notice that in the auxiliary system (16) one could also condition the probability of each indicator to the controls $\boldsymbol{x}$, in which case the unobservable types are defined relatively to $\boldsymbol{x}$. For example, if Z_h is the choice of wearing seat belts which acts as an indicator of risk preference, conditioning say on age, helps identifying risk attitudes relatively to age, while if no conditioning is made, one tends to identify unadjusted risk preferences. In our experience, for the purpose of estimating the parameters of the main system of interest, in practice there is little difference in the results obtained under the two approaches.

[12]In fact, since the α's are the only model parameters after conditioning to $\boldsymbol{x}$, and since the α's are one-to-one and differentiable functions (i.e. reparameterizations) of the probabilities of interest, model

The multivariate nature of our model is key for linking the unobservable determinants of demand and the unobserved determinants of costs, using identifying variation from the auxiliary set of indicators to uncover residual heterogeneity.[13]

Suppose we have independent observations $(I_i, L_i, Z_{1,i}, \ldots, Z_{H,i}, \boldsymbol{x}_i)$ for a sample of N units, and let the binary variable $U_{t,i}$ indicate whether subject i belongs to type t. If $U_{t,i}$ were observable, the complete-data log likelihood for model (14-15-16-17) could be written as

$$\Lambda = \sum_{i=1}^{N} \sum_{t=1}^{M} U_{t,i} \log P(T=t) + \sum_{i=1}^{N} \sum_{t=1}^{M} \sum_{j=1}^{S} U_{t,i} Y_{j,i} \log P(Y_{j,i}=1 \mid T=t, \boldsymbol{x}_i)$$

where, as in equation (10), we have renamed the response variables as $Y_1, \ldots, Y_S$, and the dependence of Λ on the model parameters is specified in the model equations (14-15-16-17).

Estimation of the model parameters by maximization of Λ can be obtained by several alternative algorithms. A robust estimation method is the EM algorithm (which we use in our applications), that is the standard approach for ML estimation of finite mixture models. The EM algorithm, though typically quite slow, has been shown (Dempster, Laird and Rubin [1977]) to converge to the maximum of the true likelihood. Given the binary nature of the observed variables, the E-step is equivalent to compute, for each subject, the posterior probability of belonging to each unobservable type. The M-step requires maximization of a multinomial likelihood.[14] It is well known that the EM algorithm may converge even if the model is not identified, a crucial issue for finite mixture models. Equation (12) however ensures that the model is identified for any value of $\boldsymbol{x}$.

(14-15-16-17) is actually *nonparametric* conditional to a given value of $\boldsymbol{x}$ (c.f. equation (10) in section 5.1).

[13]For comparison, the standard approach of estimating demand and cost with two separate finite mixture logit models, using the auxiliary set of indicators as controls in the types' membership probabilities, has some major shortcomings: it does not allow a direct link between the unobservable determinants of demand and expected costs; it cannot be used in the nonparametric case; and in our experience even with a rich set of covariates $\boldsymbol{x}$ generally implies a much more fragile identification of the unobserved types.

[14]We are grateful to Antonio Forcina for kindly providing the Matlab code for the EM estimation. Our estimation has been done with a code rewritten in Stata, which is available upon request.

Despite the usefulness of finite mixture models to detect underlying residual heterogeneity, one unresolved issue is how to determine the number of unobserved types M. The currently preferred approach suggests the use of Schwartz's Bayesian Information Criterion (BIC) to guide this choice, which in certain conditions is known to be consistent and generally helps preventing overparameterization (see McLachlan and Peel [2000] for a thorough introduction to finite mixture models and a review of existing criteria for the choice of the number of types). BIC is calculated from the maximized log-likelihood $L(\boldsymbol{\psi})$ by penalizing parameters' proliferation, $BIC(\boldsymbol{\psi}) = -2L(\boldsymbol{\psi}) + \upsilon log(n)$, where n denotes sample size and υ the number of parameters; the model with the lowest BIC is preferred.

5.4. **Posterior type probabilities.** An output of our procedure is an estimate of the joint distribution of the set of indicators $\boldsymbol{Z} = [Z_1, \ldots, Z_H]$ and the residual heterogeneity T. From the estimated joint distribution of $(\boldsymbol{Z}, T)$ we can get an estimate of the so called *posterior type probabilities* for some focal observable individual behaviour. Given a vector $\tilde{\boldsymbol{Z}}$ of observable indicators of focal interest, from $P(T, \tilde{\boldsymbol{Z}})$ posterior type probabilities are obtained using

$$P(T = t \mid \tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}) = \frac{P(T = t, \tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}})}{P(\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}})}.$$

Posterior type probabilities help understanding the underlying structure of residual heterogeneity, since they may give useful information on the nature of the estimated types, and are key for the implementation of $\boldsymbol{z}$-conditional allocations.

5.5. **Multiple losses.** In many circumstances the insurance contract protects against multiple losses. For example, Medigap protects against high out of pocket expenses for several health care services, such as inpatient, outpatient and specialist visits. The framework above can then be extended by simultaneously considering say J binary outcomes L_j, $j = 1, \ldots, J$, which take value one if the individual experiences the loss of type j for which he is insured. The demand and cost functions of interest are

$$\begin{aligned} &P(I=1 \mid \boldsymbol{x}, t)=F(\alpha_I(t)+\boldsymbol{x}'\boldsymbol{\beta}_I) \\ &P(L_j=1 \mid \boldsymbol{x}, t)=F(\alpha_{L_j}(t)+\boldsymbol{x}'\boldsymbol{\beta}_{L_j}),\ j=1,\ldots,J \end{aligned} \tag{18}$$

For sharper identification of the mixture components, this system of equations of main interest is integrated by the auxiliary system, and the complete model is (16)-(17)-(18).

6. Application to the US long-term care insurance market

In a seminal paper Finkelstein and McGarry [2006] study the long-term care (LTC) insurance market in the USA. LTC expenditure risk is one the greatest financial risks faced by the elderly in the US; to get a quantitative feeling of its importance, the amount of expenditure in nursing home care in 2004 was about 1.2% of the US GDP. Finkelstein and McGarry notice that in the sample there is *negative* correlation between insurance purchase and nursing home use. They show that individuals who exhibit more cautious behavior - as measured either by their investment in preventive health care or by seat belt use - are both more likely to have long-term care insurance coverage and less likely to use long-term care, so they conclude that the market is advantageously selected.

6.1. **Data and variables definition.** We use Finkelstein and McGarry's dataset as reported in table 4 of their paper. This is a subsample of individuals in the top quartile of the wealth and income distribution without any health characteristics that might make them ineligible for long-term care insurance.[15] We use as insurance purchase and risk occurrence two binary variables, namely *LTC Insurance* which takes value one if the individual has long-term care insurance, and *Nursing Home* which takes value one if the individual enters a nursing home in the following 5 years. As observed characteristics used by insurance company ($\boldsymbol{x}$) we use the individual probability of entering a nursing home, which is calculated by Finkelstein and McGarry from a standard actuarial model, with a cubic specification.

[15]Finkelstein and McGarry's dataset is available in the AER website. We thank the authors and the AEA for their policy of providing data for published articles.

As indicators for the residual unobserved heterogeneity we use the following binary variables: *Seat Belt* which takes value 1 if the subject always wears seat belts; *No Smoking and Drinking* which takes value 1 if the subject currently does not smoke and has less than three drinks per day; *Optimism* which takes value 1 if the individual self-reported probability of nursing home utilization is higher than the insurance company estimated probability;[16] *Preventive Care* which takes value 1 if the subject has taken more gender appropriate preventive care procedures in the past year than the median value. In this sample there are 1491 individuals; about 17% have long-term care insurance in 1995 and 10% enter a nursing home in the following 5 years period.

6.2. **Results.** The finite mixture model (14-15-16-17) is estimated using the 4 indicators above to set the auxiliary system (16). To choose the number of mixture types we use Schwartz's BIC, which achieves the minimum value with two mixture types: BIC with $M = 2$ is 8656.58 and with $M = 3$ is 8676.19.

For completeness, estimates of α's and β's and their standard errors are reported in Appendix C, but for economy of space estimated coefficients are not discussed in the main text. About 30% of individuals are of type 1, and 70% of type 2. We first test for the homogeneity of σ_t by estimating the unrestricted model as explained in section 5 above. The LR is equal to 2.977, which is distributed as chi-square with 3 d.f. and has a p-value of .395. Thus, the homogeneity null in this sample cannot be rejected.

Table 1 reports the conditional probabilities by types for the six observed variables.[17] Conditionally on $\boldsymbol{x}$ there seems to be a substantial difference in insurance purchase and nursing home use between the two types; type 1 individuals are four times less likely to buy LTC insurance, but more than twice as likely to use a nursing home as types 2. The table confirms the presence of unpriced heterogeneity, and in particular of advantageous selection, since claims and coverages are anticomonotone across types. There seems to be a natural ordering of the types in terms of their cautiousness. Types 1 show a greater probability of using seat belts and preventive care, of refraining from

[16]Spinnewijn [2013] contains an interesting analysis of the role of optimism in insurance markets.
[17]The conditional probability for nursing home use and LTC insurance are averaged across insurance risk classification.

smoking and drinking, and believing that they may need a nursing home in the near future compared with insurer's prediction.

Table 1 about here

Our analysis suggests that in LTC insurance there are two unobserved synthetic markets: one for the cautious types, with a high demand curve and a low expected MC curve; and one for the reckless types, with low demand curve and a high MC curve. Before appraising the welfare implications of the resulting unpriced heterogeneity, we need to check that, within each synthetic market, claims and coverages are independent as in equation (1). As discussed in section 3.3 above, a test of the absence of residual correlation can be performed by estimating the log-odds ratios for the association between *LTC Insurance* and *Nursing Home*. The LR test statistic of their significance is equal to 0.036 with 2 d.f., with p-value .98, so we can safely conclude that, within each synthetic market, there is no significant residual correlation between individuals' valuation of the contract and expected costs.

The existence of two synthetic markets with large differences in cost structures implies potentially serious inefficiency in actual contractual practices. To calculate our welfare measures, we first need an external estimate of η. In two recent papers, Courtemanche and He [2009] and Goda [2011] find a substantial price responsiveness of the demand for LTC insurance at the current low ownership rate, with equilibrium price elasticity η estimated at -3.9 and -3.3 respectively. We use $\eta = -3.5$ for our calculations.

Using the consumer surplus formula (22), we calculate the welfare cost measures under counterfactual efficient prices in the two markets:

Table 2 about here

The welfare loss from unpriced heterogeneity is 6.6% of total costs.[18] Comparing the welfare loss for low risk types $T = 1$ with the welfare gain for the high risk types $T = 2$, a one dollar gain for the riskier types costs a whopping \$4.1 dollars to the low risks. This latest huge number can be explained by the very low value that 'reckless' type in this sample give to the insurance contract compared to the 'cautious' ones.

For robustness, we redo these calculations using a lower value for the unusually high external estimate of η. This is particularly relevant since, in line with Finkelstein and McGarry [2006], for homogeneity we have used the subsample of healthy and wealthy individuals, and there is no guarantee that the true price elasticity η in our estimated market reflects accurately those calculated using a more representative sample of LTC insurance. Using a more conservative estimate of $\eta = -2$, welfare loss is equal to 4.9% of total costs, and the relative dollar cost of the cross subsidization is \$2.5.

The aggregate market is depicted in Figure 2. The constrained-efficient uniform price ($\bar{p}$ where the MC curve crosses the demand curve) is only marginally higher than the average cost price p^*; less than 0.1% of the current welfare loss is attributable to the difference in welfare maximizing uniform price. Thus, almost all of the welfare loss of unpriced heterogeneity is due to lack of price differentiation of the two types.

Figure 2 about here

The welfare loss of the $\boldsymbol{z}$-conditional allocation using our full set of indicators – compared to the efficient allocation– is 5.1% of total cost. To implement the $\boldsymbol{z}$-conditional allocation using this particular set of observable indicators (namely seat belts, optimism, preventive care and smoking and drinking), these characteristics need to be available to insurers (or regulators) and to be immutable, which clearly does not hold in this case.

[18]We also calculate approximate welfare loss using linear demands as in (5). Welfare loss with linear demand is 8% total insurers' costs.

To sum up our findings, it seems that in the LTC insurance market there is a significant welfare loss in the market allocation due to the unpricing of 'cautiousness', which significantly affects both demand for insurance and expected costs. Pricing by cautiousness not only would cause a significant welfare improvement, but would *increase* LTC insurance purchase by almost 32% (if types were priced at their MC, the total quantity of insurance purchase would increase from 17% to 22%, since cautious individuals would substantially increase their demand for insurance, while reckless ones would decrease their demand by a much smaller amount). Since the small size of the LTC insurance market is currently an important policy concern (e.g. Brown and Finkelstein [2011]), our results suggest that targeting cautiousness could be an important element of policy intervention. Finding observable and immutable cautiousness indicators (not belonging to current pricing variables $\boldsymbol{x}$) is a worthy challenge for increasing the efficiency of the LTCI market.

7. Application to the US Medigap insurance market

A Medigap insurance plan is a health insurance contract sold by a private company to fill 'gaps' in coverage of the basic Medicare plan. Medigap plans offer additional services and help beneficiaries pay health care costs (deductibles and co-payment) that the original Medicare plan does not cover. As a relevant example for our application, Medicare's coinsurance or copayments for hospital stays, physician visits or outpatient care are covered by the Medigap plan.

The Medigap insurance market is quite interesting to study as a further application of our methods since, contrary to the LTC insurance market, it is highly regulated.[19] In a recent influential paper, Fang, Keane and Silverman [2008] consider private information in the Medigap market using data obtained by imputing HRS observations for year

[19]Federal Law affects the Medigap market at least in three ways. First, Medigap plans are standardized into ten plans,'A' through 'J', and the basic plan 'A' must be offered if any other more generous plan is also offered. Second, there is a six-month open enrollment period when people turn 65 years old where Medigap cannot refuse any person even if there are pre-existing conditions. Third, pricing criteria are mainly based on individual's age, sex, but premiums could change according to the State of residence and individuals may receive a discount according to their smoking status. Therefore, insurers are not free to offer any insurance contract at any price they choose.

2002 in the Medicare Current Beneficiary Survey. They find a negative correlation between Medigap supplemental coverage and ex post medical expenditure, and argue that individual cognitive ability is the main source driving advantageous selection in the Medigap market.

7.1. **Data and variables definition.** The HRS is a biennial survey targeting elderly Americans over the age of 50 sponsored by the National Institute on Aging. Although the survey is not conducted on an yearly basis, from 1992 it provides longitudinal data for an array of information, consistently administrated, on several different fields such as health and health care utilization, type of insurance coverage.

We use data from waves 2-9 of the HRS, covering from 1994 to 2010. Since the interview is conducted every two years and the a six-month open enrollment period starts when people turn 65 years old, we focus on the coverage decision for the individuals who are either 65 or 66 years old at the time they were interviewed. Moreover, as mentioned before, since the structure of the contract depends not only by individual age, but also to some other specific characteristics, the pricing vector $\boldsymbol{x}$ contains the following dummies: *gender* equals 1 if individual is a female, *Smoking Status* equals 1 if individual currently smokes and then nine variables one for each of the following census division of residence: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain and Pacific.[20]

Following Fang, Keane and Silverman [2008] we define Medigap status (*Medigap*) to be equal to one if an individual is covered by Medicare and has deliberately purchased a supplemental plan additional to Medicare. We excluded from the dataset individuals who were younger than 65 years and older then 67 in each wave, or are also enrolled in any other public program or receive Medigap insurance coverage by his/her or spouse's former employer. Claims are measured by the following binary variables which take 1 if an individual: i) had any hospital stay (*Hospital*); ii) had a number of doctor visits

[20]We use the census division of residence rather than individual state of residence since this is the most detailed information on individual place of residence available in the HRS.

greater than the median (*Doctor*); iii) used any outpatient service such as surgery or home care facilities in the twelve months prior to the interview (*Out. Services*). As additional indicators to identify unobserved types we also use *Subjective Health*, which equals 1 if the individual reports good or very good health; *Income* which equals 1 if the individual is in the top wealth quartile; *Hospital-1*, *Doctor-1*, *Out. Services-1* which indicate whether the individual used medical care in the previous wave; *Insurance-1* which indicates if individual was covered by a health insurance in the previous wave; *Life Insurance* which equals 1 if the individual has a life insurance; *Annuity* which indicates if the individual purchased a life annuity; *Safe Assets* which equals 1 if the individual has a share of portfolios invested in Treasury bills or savings accounts greater than those invested in stock; *Job Number* equals 1 if individual had a number of jobs lower than the median during his/her job history; *Subjective Probability of Nursing Home Use* which equals 1 if the individual reports a probability of moving to a nursing home in next 5 years greater than the median. There are 5432 observations in the sample, 55% of females and 14% of smokers.[21]

7.2. **Results.** We estimate the finite mixture model (16)-(17)-(18) using the 11 indicators discussed above to set the auxiliary system (16). Table 3 reports the maximized log-likelihood and the BIC for different numbers of unobserved types.

Table 3 about here

The BIC seems to indicate that five types are adequate to represent any residual unobserved heterogeneity conditional on pricing variables $\boldsymbol{x}$. Therefore we report and comment the main results only of the model with the better fit.[22]

[21]Following Fang, Keane and Silverman [2008], we have also considered cognitive skills as indicator. However, variables measuring cognitive skills are available only for a single wave, so including it in the analysis would drastically reduce the sample size.

[22]Estimated parameters for the other cases are available from the authors. The general picture emerging in other cases is similar to the results discussed below. For completeness we report estimates of α's and their standard errors in Appendix C, but for economy of space estimated coefficients are not discussed in the main text.

Table 4 reports the estimated conditional probabilities by types for the four variables of interest and the eleven auxiliary indicators.

Table 4 about here

To calculate the expected marginal costs for the five types, since there are multiple losses, we use an estimation of the relative proportion of the cost of claims in the three medical care use variables. According to a study conducted in 2006 for America Health Insurance Plans (AHIP) (PriceWaterhouseCoopers [2006]), about 18% of health care insurance premiums cover inpatient care, 24% physician services and 22% outpatient care.[23] After weighting the three medical care use variables by their relative weight, we find that expected marginal costs for the five types is equal to $c_1 = 0.234$, $c_2 = 0.358$, $c_3 = 0.348$, $c_4 = 0.071$ and $c_5 = 0.677$; and estimated expected average cost is 0.298.[24]

Expected cost estimates indicate a striking heterogeneity in claims across different types. Types 2, 3 and 5 are high risk, and are on average 5 times more likely to use medical resources than low risk types 1 and 4. Looking at coverage probabilities across types we see that claims and coverages are not comonotone; for example, low risk types 2 have a much lower probability of buying insurance compared to high risk types 3 and 5, but the opposite is true for low risk types 4. Lack of comonotonicity between claims and coverages suggests multidimensional heterogeneity. This is confirmed also by looking at Panel B in Tables 4, where it emerges that there is no clear unique common underlying order of the types in terms of wealth, subjective health, past use of medical resources and insurance coverage, and risk preferences.

7.3. **Welfare.** We first test for absence of residual loss/coverage correlation within each synthetic market by estimating 15 log-odds ratios for the association of *Medigap*

[23]The remaining portion going to prescription drugs, home and nursing home care, government payments and administrative costs, consumer services and profits.

[24]Since what matters in our approach is to capture differences in expected costs between types rather than between different claims, it is unlikely that this approach introduces significant distortions in the results.

with each of the three medical care variables (three log-odds ratios for each of the five markets). The LR test statistic is equal to 16.95 with p-value 0.3216. We can safely conclude that, within each unobserved market, there is no significant residual correlation between individuals' valuation of the contract and expected costs, which is a necessary condition for valid policy analysis.[25]

We recover the parameters needed for our policy analysis from the estimated parameters in Table 4, coupled with external information on the equilibrium price elasticity of the demand for Medigap insurance, which has been recently estimated by Starc [2010] at -1.11. Using the consumer surplus formula (22), we can calculate social welfare under counterfactual efficient prices in the five markets.

Before calculating the inefficiency of cross-subsidization in the five synthetic markets, we use our estimates to examine the characteristics of the *aggregate* Medigap market implied by our estimates. Figure 3 illustrates the aggregate market.

Figure 3 about here

In the aggregate market, the difference between the AC and MC curves is very small. The constrained efficient social welfare W^{ce} is almost identical to social welfare in the market allocation W^m, and the aggregate MC curve test suggests absence of selection. This is due to the fact that aggregate willingness to pay does not map into a unique value of individuals' riskiness, and average realized costs come from a finite mixture of high and low risk individuals, so that the aggregate MC curve is almost flat.

The five synthetic markets show that there is a large difference in expected costs between high and low risk individuals, with current contracts causing substantial inefficiencies due to the implicit cross-subsidization of high risks at the expense of low risks. The social welfare of the market allocation is 75% of the efficient social welfare; the welfare loss of the market allocation as a proportion of total coverage costs ($\frac{W^e - W^m}{TC(p^*)}$)

[25]We also tested the assumption that σ_t is constant across synthetic markets. The LR test statistic is equal to 54.31 with p-value 0.0651. Thus we do not reject the null hypothesis of σ homogeneity on our policy estimates.

is 24.5%.[26] One dollar subsidy to the high risk types 2, 3 and 5 costs \$2.6 to the low risk types 1 and 4.

The $\boldsymbol{z}$-conditional allocation using past health care utilization indicators (namely *Hospital-1*, *Doctor-1* and *Out. Services-1*) allows to reduce welfare loss from 24.5% to 12.7% as a proportion of total cost, and to capture 86.8% of the efficient social welfare. The table below summarizes the relative efficiency of the various allocations.

Table 5 about here

7.4. **Robustness.** For robustness, we again redo these calculations: *i*) using 2/3 and 4/3 of Starc's η estimate, and *ii*) using $\eta = -1.11$ but assuming a different number of synthetic markets.

i) When $\eta = -0.74$, social welfare in the market allocation is 86% of the efficient social welfare, the welfare loss scaled to total costs is 17.3%, and the relative dollar cost of the cross subsidization is \$1.9. When $\eta = -1.48$, market welfare is 64% of the efficient one, and welfare loss scaled to total costs is 30.8%. The relative dollar cost of the cross subsidization is \$3.28.

ii) Table 6 below shows some estimated efficiency measures for various numbers of synthetic markets. A glance at the table shows that estimates are quite robust to the number of synthetic markets used in the analysis.

Table 6 about here

7.5. **Implications.** In the Medigap market we find multidimensional unpriced heterogeneity, with large differences across types in claims and attitudes to buy insurance.

[26]Approximate welfare loss using (5) is 23.5% of total insurers' costs.

Such large heterogeneity implies the existence of both adverse and advantageous selection in different segments of the population, and very large inefficiency of the current market allocation compared to the socially efficient one.

From the policy perspective, we find that the constrained efficient allocations which imposes uniform prices does not lead to any significant increase in efficiency as compared to the market allocation. The inefficiency of the Medigap insurance markets stems entirely from the lack of price differentiation across different types, and regulation preventing categorical discrimination seems to be a rather inefficient way to subsidize individuals with high health risk.

Allowing insurers to price individual contracts using additional information such as past health care utilization would result into a substantial efficiency increase. While the magnitude of the possible efficiency gains of non-uniform pricing is quite dependent on the true price elasticity of the demand for Medigap insurance, all plausible scenarios suggest that these efficiency gains can be large, and reforming the Medigap market may give substantial returns.

Of course these are counterfactuals based on a simple static model which ignores dynamic issues such as risk reclassification.[27] Moreover this approach assumes that the marginal utility of money is constant between types, and does not consider equity issues. However it does point out that the current heavily regulated situation has a large potential for improvement. The design of an efficient pricing scheme to uncover differences in expected costs, however, involves subtle dynamic issues and is clearly outside the scope of the present paper.

[27]See Handel, Hendel and Whinston [2013] for a recent analysis of reclassification risk.

8. Conclusion

In this paper we considered the welfare loss of unpriced heterogeneity in insurance markets, which results when private information or regulatory constraints prevent insurance companies from setting premiums reflecting expected costs.[28] The main contribution of the paper is:

(*i*) we uncover structural heterogeneity in claims and coverage, segment the market into types which differ systematically in preferences and risk, and show how the assumption that within each synthetic market all heterogeneity is idiosyncratic can be tested;

(*ii*) we explore how non-uniform pricing can improve social welfare by pricing each synthetic market appropriately. Using external information of the price elasticity of the aggregate demand for insurance, we quantify the welfare gain of efficient uniform and non-uniform prices, explore the possibility of pricing on observables not currently used in the market, and quantify the costs of cross-subsidization in a transparent way;

(*iii*) we show in our applications that uncovering all structural heterogeneity is key for a realistic policy appraisal of insurance markets.

We apply our methods to the US LTC and Medigap insurance markets, where we find that unpriced heterogeneity causes substantial inefficiency, due to the implicit cross-subsidization of high risks at the expense of low ones. The quantitative welfare loss measures we find are generally higher than existing ones (e.g. Carlin and Town [2009], Einav, Finkelstein and Cullen [2010] and Bundorf, Levin and Mahoney [2012]). In doing these comparisons, it should be however kept in mind that our results refer to specific markets, which differ from the market for employers-sponsored health plans considered by these papers. Furthermore, the welfare loss measure in Einav, Finkelstein and Cullen [2010] and in Carlin and Town [2009] (the former reports an estimate of relative welfare loss equal to 2%; the latter lower than 1%) refer to the inefficiency of

[28] There may be other reasons which prevent setting premiums equal to expected costs. For example, political economy considerations may prevent setting premiums according to, say, sexual orientation, even if expected costs depended on it. We have been told that when insurance companies tried to use some zip codes as proxies for sexual orientation, they were eventually prohibited to do so by regulators.

current pricing compared to optimally set uniform prices. Bundorf, Levin and Mahoney [2012] estimated a welfare loss of 1-13% of total coverage costs, comparing current prices with risk-rated premiums, which is more in line with our approach.

Our larger estimated welfare loss can probably be explained in the LTCI market by the large price elasticity of the demand for insurance, and in the Medigap market by the rather extreme regulation, particularly in the open enrollment period. Our results seem to be fairly robust to different specifications, and most of the action seems to come from the very large differences in expected costs that we extract between unobserved types.

The methodology discussed in this paper allows using large representative samples, and given the wealth of commonly available and well known survey data, allows exploring the efficiency of many insurance markets where the researcher does not have access to firm data. Using detailed individual information from survey data also helps the extraction of systematic differences in insurance preferences and expected claims. The most relevant limitation of our approach is probably the need to calibrate the model with external information on the price elasticity of the aggregate demand for insurance.

The proposed methodology is not necessarily limited to survey data. If the researcher had access to appropriate administrative or firm data with exogenous variation in insurance premiums, there would be no need for external elasticity information. Finding rich sets of individual indicators of risk and preferences may be more challenging with firm or administrative data.

9. Table and Figures

Table 1. Estimated conditional probabilities by types

	T=1	T=2
Seat Belt	0.6114	0.9323
Optimism	0.3546	0.4912
Preventive Care	0.2541	0.4491
No Smoking and Drinking	0.7904	0.9492
Long-Term Care Insurance	0.0733	0.2875
Nursing Home	0.3804	0.1978

Table 2. Welfare Measures ($j = e, z, ce, m$)

	$\frac{W^e - W^j}{W^e}$	$\frac{W^e - W^j}{TC(p^*}$
W^e	0	0
W^z	16.3%	5.1%
W^{ce}	21.0%	6.5%
W^m	21.2%	6.6%

Table 3. Log-likelihood and BIC

	2LC	3LC	4LC	5LC	6LC
Log-Likelihood	-47092.89	-46827.48	-46662.6	-46554.64	-46491.92
BIC	94796.38	94403.16	94211.02	94132.7	94144.85
# of Parameters	71	87	103	119	135

Table 4. Estimated conditional probabilities

	T=1	T=2	T=3	T=4	T=5
Panel A: Main Equations					
Hospital	.1571	.1935	.2224	.0428	.6066
Doctor	.3907	.5367	.5215	.0417	.8973
Out. Services	.1248	.2975	.2610	.1240	.4939
Medigap	.1107	.0857	.6686	.3035	.2342
Panel B: Auxiliary Indicators					
Hospital-1	.1624	.1718	.1819	.0316	.4492
Doctor-1	.4388	.6297	.5703	.0611	.8513
Out. Services-1	.1083	.2856	.2161	.0805	.3448
Insurance-1	.0529	.0177	.4940	.2003	.1036
Sub. Health	.0300	.4749	.2994	.2610	.1239
Life Insurance	.2662	.5624	.4481	.7529	.0653
Annuity	.6534	.7903	.6013	.7056	.7095
Income	.1628	.7803	.7072	.6191	.3975
Safe Assets	.0930	.4119	.3623	.3448	.1785
Job Number	.6181	.5458	.5951	.5969	.6866
Sub. Prob. Nursing Home Use	.3695	.5863	.5748	.4634	.5002
Panel C: Types Proportion					
	0.21	0.25	0.13	0.26	0.15

TABLE 5. Welfare Measures ($j = e, z, ce, m$)

	$\frac{W^j}{W^e}$	$\frac{W^e - W^j}{TC(p^*)}$
W^e	1	0
W^z	86.8%	12.7%
W^{ce}	75.0%	24.5%
W^m	74.9%	24.5%

TABLE 6. Welfare Measures With Different Number of Synthetic Market

M	$\frac{W^m}{W^e}$	$\frac{W^m - W^e}{TC(p^*)}$	Cross-Sub
2	78.1%	21.9%	\$2.08
3	78.4%	21.4%	\$2.03
4	76.4%	23.6%	\$2.22
5	74.9%	24.5%	\$2.60
6	74.7%	24.9%	\$2.48

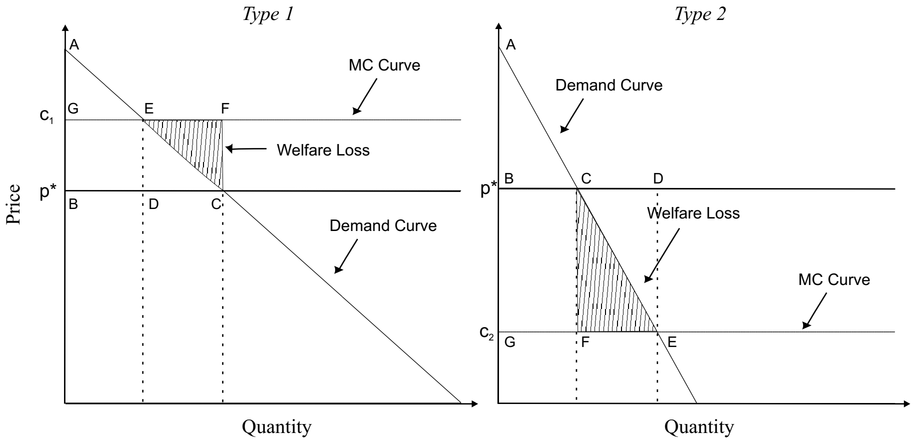

FIGURE 1. Synthetic markets for types 1 and 2

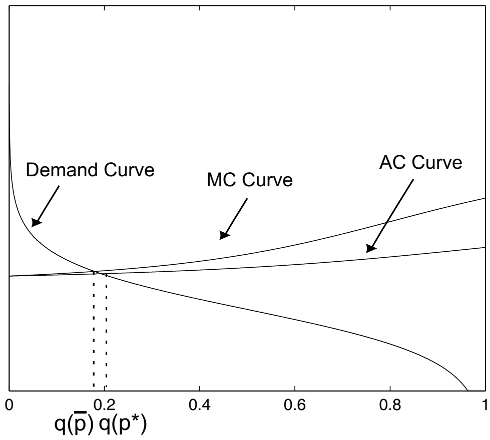

FIGURE 2. LTC: Aggregate market

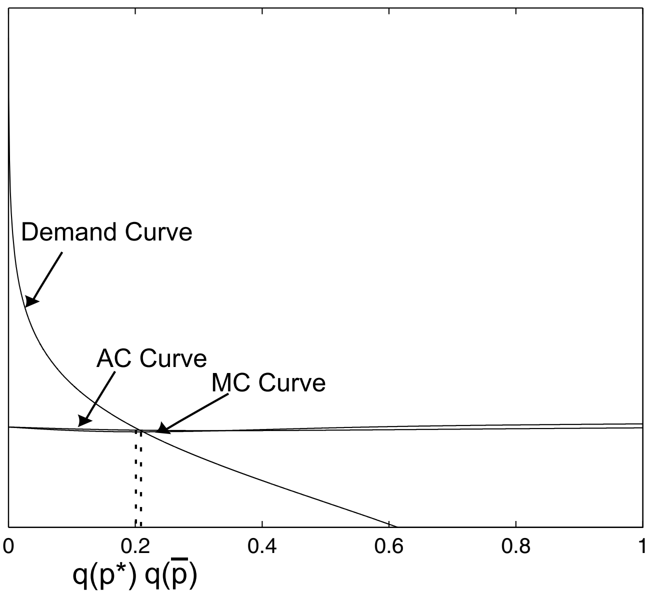

FIGURE 3. Medigap: Aggregate market

References

Allman, Elizabeth S., Catherine Matias and John A. Rhodes. 2009. "Identifiability of Parameters in Latent Structure Models with many Observed Variables." *The Annals of Statistics* 37(6A):3099–3132.

Brown, Jeffrey R. and Amy Finkelstein. 2011. "Insuring Long-Term Care in the United States." *Journal of Economic Perspectives* 25(4):119–42.

Bundorf, M. Kate, Jonathan Levin and Neale Mahoney. 2012. "Pricing and Welfare in Health Plan Choice." *American Economic Review* 102(7):3214–48.

Carlin, Caroline and Robert Town. 2009. "Adverse Selection, Welfare and the Optimal Pricing of Employer Sponsored Health Plans." *Manuscript. U. of Minnesota* .

Courtemanche, Charles and Daifeng He. 2009. "Tax Incentives and the Decision to Purchase Long-Term Care Insurance." *Journal of Public Economics* 93(1-2):296–310.

Cutler, David M., Amy Finkelstein and Kathleen McGarry. 2008. "Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance." *American Economic Review* 98(2):157–162.

Dardanoni, Valentino and Paolo Li Donni. 2012. "Incentive and Selection Effects of Medigap Insurance on Inpatient Care." *Journal of Health Economics* 31(3):457 – 470.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

Einav, Liran and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25(1):115–38.

Einav, Liran, Amy Finkelstein and Jonathan Levin. 2010. "Beyond Testing: Empirical Models of Insurance Markets." *Annual Review of Economics* 2(1):311–336.

Einav, Liran, Amy Finkelstein and Mark R. Cullen. 2010. "Estimating Welfare in Insurance Markets Using Variation in Prices." *The Quarterly Journal of Economics* 125:877–921(45).

Fang, Hanming, Michael P. Keane and Dan Silverman. 2008. "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market." *Journal of Political Economy* 116(2):303–350.

Finkelstein, Amy, Erzo F. P. Luttmer and Matthew J. Notowidigdo. 2009. "Approaches to Estimating the Health State Dependence of the Utility Function." *American Economic Review* 99(2):116–21.

Finkelstein, Amy and James Poterba. 2006. Testing for Adverse Selection with "Unused Observables". Working Paper 12112 NBER.

Finkelstein, Amy and Kathleen McGarry. 2006. "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market." *The American Economic Review* 96(4):938–958.

Follmann, Dean A. and Diane Lambert. 1989. "Generalizing Logistic Regression by Nonparametric Mixing." *Journal of the American Statistical Association* 84(405):295–300.

Geruso, Michael. 2012. "Selection in Employer Health Plans:Homogeneous Prices and Heterogeneous Preferences." *Manuscript. Princeton University* .

Goda, Gopi Shah. 2011. "The Impact of State Tax Subsidies for Private Long-Term Care Insurance on Coverage and Medicaid Expenditures." *Journal of Public Economics* 95(7-8):744–757.

Handel, Benjamin R., Igal Hendel and Michael D. Whinston. 2013. Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk. Working Paper 19399 National Bureau of Economic Research.

Heckman, James and Burton Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52(2):271–320.

Joreskog, Karl G. and Arthur S. Goldberger. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association* 70(351):631–639.

Lustig, Joshua. 2011. "Measuring Welfare Losses from Adverse Selection and Imperfect Competition in Privatized Medicare." *Manuscript. Boston University* .

Matzkin, Rosa L. 2007. "Nonparametric Identification." *Handbook of Econometrics* 6:5307–5368.

McLachlan, Geoffrey J. and David Peel. 2000. *Finite Mixture Models*. Wiley-Interscience.

PriceWaterhouseCoopers. 2006. *The Factors Fueling Rising Healthcare Costs 2006*. Report for the America's Health Insurance Plans (AHIP).

Rothschild, Michael and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *The Quarterly Journal of Economics* 90(4):629–649.

Small, Kenneth A and Harvey S Rosen. 1981. "Applied Welfare Economics with Discrete Choice Models." *Econometrica* 49(1):105–30.

Smart, Michael. 2000. "Competitive Insurance Markets with Two Unobservables." *International Economic Review* 41(1):153–69.

Spinnewijn, Johannes. 2013. "Insurance and Perceptions: How to Screen Optimists and Pessimists." *The Economic Journal* 123(569):606–633.

Starc, Amanda. 2010. "Insurer Pricing and Consumer Welfare: Evidence from Medigap." *Manuscript. Harvard University* .

Suppes, Patrick and Mario Zanotti. 1981. "When are probabilistic explanations possible?" *Synthese* 48(2):191–199.

Wooldridge, James M. 2002. *Econometric analysis of cross section and panel data*. The MIT press.

Appendix A. Logit demand

Using the logit utility model (6), the demand function in market t is given by

$$q_t(p) = \frac{\exp\big(-\frac{p-w_t}{\sigma}\big)}{1+\exp\big(-\frac{p-w_t}{\sigma}\big)}. \tag{19}$$

Without further information we cannot identify w_t and σ in equation (19). Our strategy is to use external information, namely the equilibrium price elasticity of the aggregate demand for insurance, to identify σ.

Differentiate then (19) w.r.t. p

$$\frac{\partial q_t(p)}{\partial p} = -\frac{1}{\sigma}(1-q_t(p))q_t(p), \tag{20}$$

so that, given aggregate demand $q(p) = \sum_t P(T=t)q_t(p)$,

$$\frac{\partial q(p)}{\partial p} = -\frac{1}{\sigma}\sum\nolimits_t P(T=t)(1-q_t(p))q_t(p).$$

Given an external estimate of the price elasticity of the aggregate demand η at the aggregate equilibrium $(p^*, q(p^*))$, we have

$$\eta = -\frac{\big(\sum_t P(T=t)(1-q_t(p^*))q_t(p^*)\big)p^*}{\sigma q(p^*)}$$

and so

$$\sigma = -\frac{\big(\sum_t P(T=t)(1-q_t(p^*))q_t(p^*)\big)p^*}{\eta q(p^*)}. \tag{21}$$

From estimated p^* and σ and equation (19) we derive w_t for each market t, which is all we need to calculate consumer surplus for t-type individuals, as given by

$$CS_t(p_t) = \sigma \log\big[1+\exp\big(-\frac{p_t-w_t}{\sigma}\big)\big], \tag{22}$$

which can be compared with the Rosen and Small [1981] multinomial logit logsum measure.

Summing up, from estimated $\big(P(T=t), P(I=1 \mid t), P(L=1 \mid t)\big)$ for each synthetic market, the average cost assumption allows an estimate of p^*, and external information on η gives an estimate of σ. This gives all the required parameters to allow quantitative welfare analysis and efficiency evaluations of counterfactual policies.

A.1. **A worked example.** Suppose an insurance market is composed of two unobserved types, say high risk $t = 1$ and low risk $t = 2$. We observe aggregate claims and insurance demand probabilities, say $P(L = 1) = .5$ and $P(I = 1) = .5$. As a first step, we need to estimate the claims and insurance demand probabilities in the two unobserved synthetic markets, and the types probabilities. Since the joint distribution of I, L has only three independent moments, without further variation we cannot estimate the five needed parameters.

Suppose we have two observed binary risk indicators, say Z_1 and Z_2 with $P(Z_1 = 1) = P(Z_2 = 1) = .5$, such that I, L, Z_1, Z_2 are independent conditional on T. The joint distribution of I, L, Z_1, Z_2 contains 16 observed independent moments which are equal to, say,

$$\begin{aligned}
P(I=1,L=1,Z_1=1,Z_2=1) &= \textstyle\sum_{t=1}^{2} P(T=t)P(I=1\mid T=t)P(L=1\mid T=t)P(Z_1=1\mid T=t)P(Z_2=1\mid T=t)\\
P(I=0,L=1,Z_1=1,Z_2=1) &= \textstyle\sum_{t=1}^{2} P(T=t)P(I=0\mid T=t)P(L=1\mid T=t)P(Z_1=1\mid T=t)P(Z_2=0\mid T=t)\\
\vdots \qquad & \vdots \qquad\qquad \vdots\\
P(I=1,L=0,Z_1=0,Z_2=0) &= \textstyle\sum_{t=1}^{2} P(T=t)P(I=1\mid T=t)P(L=0\mid T=t)P(Z_1=0\mid T=t)P(Z_2=0\mid T=t).
\end{aligned} \tag{23}$$

This system can then be written more compactly as $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{q}$ is the 16×1 vector of observed moments which arrays the joint distribution of I, L, Z_1, Z_2, while $\boldsymbol{\theta}$ denotes the 1×9 vector of unknown parameters (the T-conditional distribution of I, L, Z_1, Z_2 and the marginal distribution of T). Suppose $\boldsymbol{q}$ is equal to (0.1524, 0.1026, 0.0276, 0.0274, 0.0676, 0.0474, 0.0324, 0.0426, 0.0426, 0.0324, 0.0474, 0.0676, 0.0274, 0.0276, 0.1026, 0.1524). Solving the following problem by nonlinear least squares:

$$min_\theta(\boldsymbol{q} - \boldsymbol{f}(\boldsymbol{\theta}))'(\boldsymbol{q} - \boldsymbol{f}(\boldsymbol{\theta}))$$

gives the 9 parameters we need for our analysis.[29]

$$
\begin{aligned}
&P(T=1)=P(T=2)=.5;\\
&P(I=1\mid T=1)=q_1(p^*)=0.8;\ P(I=1\mid T=2)=q_2(p^*)=0.2;\\
&P(L=1\mid T=1)=c_1=0.7;\ P(L=1\mid T=2)=c_2=0.3;\\
&P(Z_1=1\mid T=1)=0.9;\ P(Z_1=1\mid T=2)=0.1;\\
&P(Z_2=1\mid T=1)=0.6;\ P(Z_2=1\mid T=2)=0.4.
\end{aligned}
\tag{24}
$$

From these, we directly calculate the average cost price $p^* = \frac{\sum_t P(T=t)q_t c_t}{\sum_t P(T=t)q_t} = 0.62$ and the aggregate quantity $q(p^*) = 0.50$. Using an external estimate of the equilibrium price elasticity of the aggregate market, say $\eta = -1$, we get $\sigma = 0.20$, $w_1 = 0.89$ and $w_2 = 0.34$. We find the efficient allocation in the two synthetic markets ($p_1 = 0.70, q_1(p_1) = 0.73$) and ($p_2 = 0.30, q_2(p_2) = 0.56$), which can be compared with the actual allocation ($p^* = 0.62, q_1(p^*) = 0.80$) and ($p^* = 0.62, q_2(p^*) = 0.20$). 50% of individuals are actually insured in this market, compared to the efficient quantity (66%).

Using the consumer surplus formula (22), we can calculate the welfare loss measures under counterfactual prices in the two markets. To find the $\boldsymbol{z}$-conditional welfare we first need to find the posterior types' probabilities: $P(T = 1|Z_1 = 0, Z_2 = 0) = 0.07$, $(P(T = 1|Z_1 = 0, Z_2 = 1) = 0.14$, $P(T = 1|Z_1 = 1, Z_2 = 0) = 0.86$, $(P(Z_1 = 1, Z_2 = 1) = 0.93$, and use these types' probabilities to compute the expected costs for each of these four groups (for example we price individuals with $Z_1 = 0$ and $Z_2 = 0$ at their expected cost $0.07 * 0.7 + 0.93 * 0.3 = 0.33$). To calculate the constrained efficient social welfare, we find the uniform price which maximizes social welfare being equal to $\bar{p} = 0.38$, which is lower than the average cost price $p^* = 0.62$

[29]It is important to notice that this is just a special theoretical example which illustrates the underlying intuition on the identification of the parameters. For estimation, as argued in section 5.3, we use the EM algorithm as it is more robust and can handle more complex models.

TABLE 7. Welfare Loss

	$\frac{W^j}{W^e}$	$\frac{W^e - W^j}{TC(p^*)}$
W^e	1	0
W^z	95.8%	2.84%
W^{ce}	92.7%	4.94%
W^m	86.7%	9.00%

The total welfare loss of the current market allocation from unpriced heterogeneity is 9% of total costs. The constrained efficient allocation welfare loss equals 4.94%, while the z-conditioned is 2.84%. One dollar gain in the market allocation for the riskier types costs about 1.91 dollars to the low risks.

A.2. **Robustness and calibration.** A major feature of our approach is that we cannot identify the scale parameter σ in the utility equation (6) and our welfare calculations depend on external calibration through η. Since σ is a relevant parameter for the calculation of consumer surplus (see equation (22)) which is at the core of our counterfactual analysis, for robustness the researcher may experiment with different values of η taken from different papers in the same market, other literature in similar markets, or simply informed guess. Furthermore, as already noticed, in equation (6) it is assumed that σ does not vary among different types t.

In this example, $\eta = -2/3$ and $\eta = -4/3$ give respectively a relative cost of cross-subsidization equal to \$1.56 and \$2.28, compared with \$1.91 when $\eta = -1$. To check the effect of heterogeneity in σ, the researcher can also experiment with different values of σ in each market. For example, we can compare welfare loss calculated under $\sigma = 0.20$ as above, with alternative scenarios such as σ_1 being twice or half σ_2. When $\sigma_1 = .15$ and $\sigma_2 = .30$, relative cost of cross-subsidization is \$1.60; when $\sigma_1 = .30$ and $\sigma_2 = .15$, it is \$2.19.[30]

As a final robustness check, we can also compare the estimates obtained under logit demand functions with the approximate deadweight loss (5). In this example, approximate welfare loss of the market allocation is about 7.1% of total cost, compared to 9%

[30]These results are specific to this example; in general there is no obvious relation between σ_t, q_t and c_t.

under the true model with logit demands. Since we do not know the exact functional form of insurance demand, equation (5) is a useful tool for checking the robustness of measured welfare loss.

Appendix B. An example with multidimensional heterogeneity

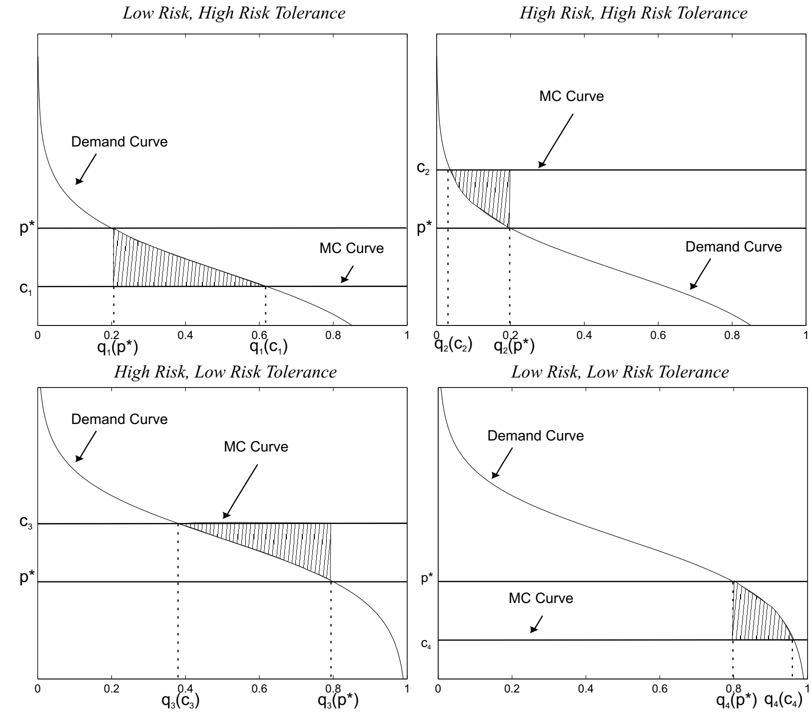

Figure 4. Synthetic markets

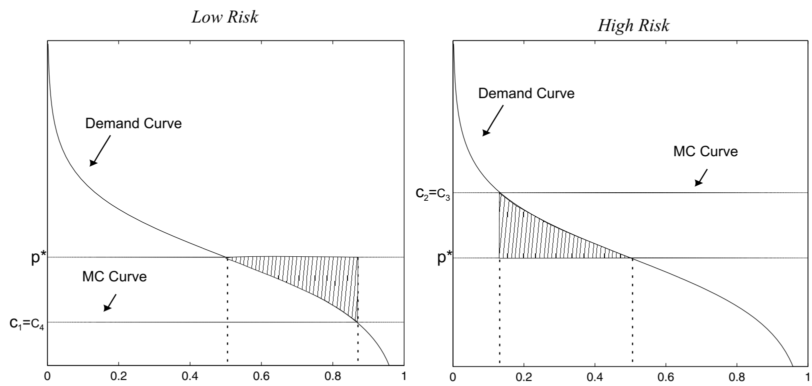

Figure 5. Synthetic markets aggregated by risk

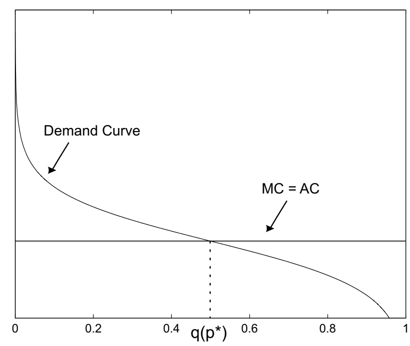

Figure 6. Aggregate market

Appendix C. Tables of estimated parameters

Table 8. Long-Term Care Insurance: Estimated α parameters

	$T = 1$	$T = 2$
Panel A: eq. 14-15		
Long-term Care Insurance	-2.867	-1.231
	(0.728)	(0.139)
Nursing Home	-1.853	-2.894
	(0.263)	(0.247)
Panel B: eq. (16)		
Seat Belt	0.453	2.623
	(0.356)	(0.478)
Optimism	-0.599	-0.0354
	(0.203)	(0.0977)
Preventive Care	-1.077	-0.204
	(0.273)	(0.113)
No Smoking and Drinking	1.327	2.929
	(0.273)	(0.361)
Panel C: eq. (17)		
	0.842	-
	(0.565)	-

Standard errors in brackets.

Table 9. Long-Term Care Insurance: Estimated β Parameters

	LTCI	NH
Risk Classification	0.926	1.645
	(0.510)	(0.627)
Risk Classification^2	-0.0531	-0.0197
	(0.0251)	(0.0274)
Risk Classification^3	0.0647	-0.000171
	(0.0309)	(0.0326)

Standard errors in brackets.

Table 10. Medigap: Estimated α parameters

	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$
Panel A: eq. (18)					
Hospital	-1.715	-1.462	-1.285	-3.146	0.412
	(0.248)	(0.234)	(0.257)	(0.314)	(0.241)
Doctor	-0.700	-0.0964	-0.159	-3.415	1.957
	(0.220)	(0.213)	(0.239)	(0.491)	(0.304)
Out. Services	-1.776	-0.685	-0.867	-1.784	0.153
	(0.231)	(0.194)	(0.221)	(0.209)	(0.202)
Insurance	-2.290	-2.580	0.595	-1.001	-1.368
	(0.268)	(0.409)	(0.340)	(0.216)	(0.231)
Panel B: eq. (16)					
Hospital-1	-1.640	-1.573	-1.504	-3.423	-0.204
	(0.118)	(0.103)	(0.154)	(0.282)	(0.0978)
Doctor-1	-0.246	0.531	0.283	-2.732	1.745
	(0.0974)	(0.0962)	(0.142)	(0.361)	(0.165)
Out. Services-1	-2.108	-0.917	-1.289	-2.436	-0.642
	(0.153)	(0.0836)	(0.142)	(0.149)	(0.0959)
Insurance-1	-2.886	-4.017	-0.0239	-1.384	-2.158
	(0.245)	(1.029)	(0.229)	(0.0881)	(0.161)
Sub. Health	-1.014	0.251	-0.208	1.114	-2.661
	(0.119)	(0.0889)	(0.130)	(0.102)	(0.297)
Life Insurance	0.634	1.327	0.411	0.874	0.893
	(0.0856)	(0.101)	(0.128)	(0.0749)	(0.0999)
Annuity	-1.638	1.268	0.882	0.486	-0.416
	(0.213)	(0.124)	(0.143)	(0.0776)	(0.101)
Income	-3.476	-0.101	-0.850	-1.041	-1.956
	(0.519)	(0.0905)	(0.134)	(0.0809)	(0.159)
Safe Assets	-2.277	-0.356	-0.566	-0.642	-1.526
	(0.199)	(0.0800)	(0.119)	(0.0729)	(0.127)
Job Number	0.481	0.184	0.385	0.393	0.784
	(0.0844)	(0.0742)	(0.116)	(0.0690)	(0.0999)
Prob. Nur. Home	-0.534	0.349	0.302	-0.147	0.000865
	(0.0895)	(0.0773)	(0.117)	(0.0688)	(0.0908)
Panel C: eq. (17)					
	0.186	-0.676	0.696	-0.503	-
	(0.128)	(0.230)	(0.197)	(0.110)	-

Standard errors in brackets.

TABLE 11. Medigap: Estimated β parameters

	Hospital	Doctor	Out. Services	Insurance
Female	-0.151	0.320	-0.0847	0.356
	(0.0782)	(0.0744)	(0.0687)	(0.0787)
Current Smoker	-0.159	-0.205	-0.0344	-0.142
	(0.113)	(0.105)	(0.0975)	(0.111)
New England	-	-	-	-
	-	-	-	-
Middle Atlantic	-0.238	0.258	-0.271	-0.535
	(0.244)	(0.221)	(0.204)	(0.235)
East North Central	0.187	0.226	-0.0457	0.367
	(0.229)	(0.211)	(0.192)	(0.216)
West North Central	0.187	-0.438	-0.285	0.679
	(0.243)	(0.226)	(0.208)	(0.227)
South Atlantic	0.210	0.218	0.0122	-0.124
	(0.223)	(0.206)	(0.187)	(0.213)
East South Central	0.363	0.344	-0.222	0.146
	(0.255)	(0.240)	(0.221)	(0.245)
West South Central	0.204	0.00176	-0.237	-0.113
	(0.245)	(0.227)	(0.209)	(0.236)
Mountain	0.0583	-0.349	-0.166	-0.324
	(0.267)	(0.248)	(0.227)	(0.261)
Pacific	0.0853	0.120	-0.137	-0.514
	(0.239)	(0.220)	(0.201)	(0.234)

Standard errors in brackets.