INFORMATION & EQUILIBRIUM IN INSURANCE MARKETS WITH BIG DATA

PETER SIEGELMAN[1]

***

Asymmetric information makes the behavior of insurance markets very difficult to predict. But this Article argues that the increasing use of Big Data by insurers will not result in forecasts of loss that are so accurate that they eliminate uncertainty, and with it, the possibility of insurance. Big Data techniques might lead to a "flip" in informational asymmetry, resulting in a situation in which insurers know more about their customers than the latter know about themselves. But the effects of such a development could actually be benign. Finally, the Article considers the potential for Big (or at least, More) Data to create new markets for spreading risks that are currently uninsurable.

***

I. INTRODUCTION

Big Data is a hot topic these days, at least in the nerdosphere.[2] Pundits proclaim it to be "revolutionary,"[3] "transformative,"[4] and "a tidal wave."[5] Some have even suggested that the further use of Big Data will overturn our outmoded reliance on primitive notions such as "causation"[6]

---

[1] I thank Peter Kochenburger, Rick Swedloff, and the editors of the CILJ for helpful comments, and Pat McCoy and Francois Ewald for initiating the conversation.

[2] VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013) (sitting at number 9,501 on Amazon.com's sales rankings as of October 13, 2014: not bad for a book with "data" in its title).

[3] *Id.* at 7.

[4] ERIC SCHMIDT & JARED COHEN, THE NEW DIGITAL AGE: TRANSFORMING NATIONS, BUSINESSES, AND OUR LIVES (2014).

[5] BILL FRANKS, TAMING THE BIG DATA TIDAL WAVE: FINDING OPPORTUNITIES IN HUGE DATA STREAMS WITH ADVANCED ANALYTICS 5 (2012).

[6] "Petabytes [lots of data] allow us to say: 'Correlation is enough.'" Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED MAG. (June 23, 2008), http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory; *see also Correlation*, XKCD, http://xkcd.com/552/ (last visited Nov. 21, 2014).

and "privacy."[7]

This Article has a much narrower focus, however: I want to reflect critically on the role of Big Data in insurance. In particular, I ask what economic theory has to say about whether Big Data will lead to new equilibria in insurance markets. I focus on three questions: Might Big Data lead to the collapse of insurance altogether by permitting predictions of such accuracy that risk and uncertainty are effectively eliminated? Even if it doesn't have such drastic effects, might it alter insurance market equilibria, by reducing the scope for risk-spreading? And might it be used to create new types of insurance that are not currently practical given current informational constraints? At the risk of destroying the narrative suspense, my proposed answers are, respectively: "no," "probably not," and "possibly."

So, what *is* Big Data, anyway? Big Data is not a precise term, and several definitions are currently competing for supremacy. For our purposes, it suffices to think of Big Data as (i) very large collections of observations, particularly those that also include very large numbers of variables;[8] and (ii) associated statistical techniques for using these ultra-large datasets to make predictions or forecasts.

II. PROLOGUE: INSURANCE MARKETS ARE WEIRD

A classic method of economic analysis is known as "Comparative Statics:" assume a (small) change to some variable, and then compare equilibria before and after this change has worked its way through the model or system. Economists have come to realize, however, that this method tends to break down in markets where there are significant informational asymmetries, that is, where one party to a transaction knows more than their counterpart does.[9] Insurance markets are the *locus classicus*

---

[7] Claire Porter, *Little Privacy in the Age of Big Data*, GUARDIAN (June 20, 2014), http://www.theguardian.com/technology/2014/jun/20/little-privacy-in-the-age-of-big-data ("In the era of big data, the battle for privacy has already been fought and lost . . . .").

[8] According to Google chief economist Hal Varian, "Google has seen 30 trillion URLs, crawls over 20 billion of those a day, and answers 100 billion search queries a month. Analyzing even one day's worth of data of this size is virtually impossible with conventional techniques." Hal R. Varian, *Big Data: New Tricks for Econometrics*, 28 J. ECON. PERSP. 3, 4 (2014).

[9] *See generally* George A. Akerlof, *The Market For "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q. J. ECON. 488 (1970) (using the used

of informational asymmetries, in the form of adverse selection and moral hazard,[10] and this in turn implies that our ordinary intuitions about how markets work may fail decisively when it comes to insurance markets.

For example, we would predict that in ordinary markets, sellers would view demand for their product as a good thing, and indeed would be delighted to sell to anyone who wanted to buy from them: picture Lucy at her lemonade stand when a customer arrives and says "I'll buy all the lemonade you have to sell at 25¢ a glass." But insurance is different. How will Irene react when someone rushes up to her insurance stand and says "I'll buy all the life insurance you'll sell me at 25¢ per $125 of coverage?" The explanation for the difference is, of course, the (fear of an) informational asymmetry that Irene faces but Lucy does not. The life insurance customer who desperately wants lots of coverage may well know something about his own prospects for longevity that her potential insurer does not know, and this information is obviously highly relevant to the insurer's profitability from transacting with this customer.[11]

It is by now well-known that informational asymmetries have a profound effect on the institutions of insurance markets, from the language of contracts to the scope and function of regulation. My point is that in the presence of such asymmetries, insurance market equilibria are highly sensitive to small and seemingly trivial details of how a market operates.

---

car market as an example to discuss the relationship between quality and uncertainty and the problem that relationship poses for the theory of market equilibrium); Kenneth J. Arrow, *Uncertainty and the Welfare Economics of Medical Care*, 53 AMER. ECON. REV. 941 (1963) (explaining that the special economic problems of the medical care industry are adaptations to the existence of uncertainty in the incidence of disease and the efficacy of treatment); Michael Rothschild & Joseph Stiglitz, *Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information*, 90 Q. J. ECON. 629 (1976) (analyzing competitive insurance markets in which the characteristics of the insured are not fully known to the insurer).

[10] Both concepts are central to virtually all aspects of modern economics; both began as terms of art in insurance. Adverse selection can loosely be defined as the tendency of the worst risks to find insurance price for an average risk to be especially attractive. Moral hazard (again, loosely) occurs whenever the presence of insurance causes insureds to take less care to prevent risks than they would exercise in its absence.

[11] *See* USLife Credit Life Ins. Co. v. McAfee, 630 P.2d 450 (Wash. Ct. App. 1981) (discussing how an insurance professional took out numerous credit life insurance policies that required no medical underwriting, on his wife, who he knew was suffering from terminal cancer).

Under some circumstances, there may be no equilibrium possible at all;[12] under slightly different circumstances, only "separating" equilibria (those in which each risk-type pays a premium that fully reflects its riskiness, with no cross-subsidization between types); under others, "pooling" (cross-subsidization from low-risk to high-risk insureds) is sustainable in equilibrium.[13] Moreover, insurance supply and demand are not actually independent, as they are in ordinary markets.[14] Thus, a mandate to buy insurance, rather than simply increasing demand and causing prices to rise, may actually lower costs and result in a *fall* in prices; it could even obviate the requirement to purchase insurance in the first instance.[15]

The situation gets even more complicated and unpredictable if we recognize that consumers are not perfectly rational, which the evidence overwhelmingly demonstrates is the case.[16] Consumers often buy "insurance" products, such as extended warranties, that no rational person should want;[17] conversely, they frequently shun coverage for losses due to floods or earthquakes that a rational person would want to insure against.[18]

---

[12] Rothschild & Stiglitz, *supra* note 9, at 634–37.

[13] *Id.*; *see also* Georges Dionne & Neil Doherty, *Adverse Selection in Insurance Markets: A Selective Survey*, *in* CONTRIBUTIONS TO INSURANCE ECONOMICS 116 (Georges Dionne ed., 1992).

[14] For a cogent explanation, see Liran Einav & Amy Finkelstein, *Selection in Insurance Markets: Theory and Empirics in Pictures*, 25 J. ECON. PERSP. 115, 118 (2011). The basic idea is that unlike a purchaser of, say, broccoli, the purchaser of insurance contributes to *both* sides of the market. A low-risk purchaser lowers the aggregate risk of the pool of insureds as a whole, and thus reduces the cost of supplying insurance to everyone. Demand and cost are not independent.

[15] Raphael Boleslavsky & Sergio J. Campos, *Does the Individual Mandate Coerce?*, 68 U. MIAMI L. REV. 1, 4-8 (2012).

[16] *See generally* HOWARD C. KUNREUTHER ET AL., INSURANCE AND BEHAVIORAL ECONOMICS: IMPROVING DECISIONS IN THE MOST MISUNDERSTOOD INDUSTRY (2013) (discussing examples of "anomalous" behavior by consumers, insurance companies, investors, and regulators).

[17] For a detailed explanation and policy recommendations, see Tom Baker & Peter Siegelman, *"You Want Insurance With That?" Using Behavioral Economics to Protect Consumers from Add-On Insurance Products*, 20 CONN. INS. L.J. 1 (2013).

[18] Tom Baker & Peter Siegelman, *Behavioral Economics and Insurance Law: The Importance of Equilibrium Analysis*, *in* OXFORD HANDBOOK OF BEHAVIORAL ECONOMICS AND THE LAW (Doron Teichman & Eyal Zamir eds., 2014); David M. Cutler & Richard Zeckhauser, *Extending the Theory to Meet the Practice of Insurance*, *in* BROOKINGS-WHARTON PAPERS ON FINANCIAL SERVICES (2004);

And it turns out that correcting for some kinds of "mistakes" made by insufficiently-rational consumers may actually exacerbate informational asymmetries and reduce welfare for everyone.[19]

The moral of all this is simple: beware of anyone (including me) who confidently tells you anything about how insurance markets behave, including how they will react to the increased use by insurers of Big Data. There is little basis in theory or empirical evidence for any confident forecast about how Big Data will shape insurance markets. What follows, then, is more by way of cautious speculation than robust prediction.

III. COULD BIG DATA VANQUISH UNCERTAINTY (AND DESTROY INSURANCE)?

A. TMI AND THE ABSENCE OF INSURANCE

Economists have long understood that uncertainty is a prerequisite for insurance. Table 1 provides a simple numerical example. A village consists of 100 identical houses, each of which is worth $200,000, and which constitutes each homeowner's total wealth. There is a 25% chance that any individual house will be completely destroyed by the next earthquake. Each homeowner has the same utility function, U_i = U(Wealth) = *ln*(Wealth), which implies that they are risk-averse.

Will the villagers demand insurance, assuming it can be purchased at the actuarially-fair premium (without any load)? To see that the answer is yes, we can compare each villager's expected utility without insurance to her utility with it. Without insurance, a homeowner's expected utility is

$$\Pr(\text{Loss})\times(\text{Utility}|\text{Loss}) + \Pr(\text{No Loss})\times(\text{Utility}|\text{No Loss}) =$$

$$0.25\times ln(\text{Wealth}|\text{Loss}) + 0.75\times ln(\text{Wealth}|\text{No Loss}) =$$

$$0.25\times ln(0) + 0.75\times ln(200{,}000) = 8.58.$$[20]

KUNREUTHER ET AL., *supra* note 16, at 115.

[19] *See, e.g.*, Benjamin R. Handel, *Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts*, 103 AMER. ECON. REV. 2643 (2013); Alvaro Sandroni & Francesco Squintani, *Overconfidence, Insurance, and Paternalism*, 97 AMER. ECON. REV. 1994, 1994 (2007); Justin Sydnor, *(Over)insuring Modest Risks*, 2 AMER. ECON. J.: APPLIED ECON. 177, 198 (2010).

[20] Since *ln*(0) is undefined, we innocuously substitute 0.001 for (Wealth|Loss).

100 times this amount is the village's aggregate utility when nobody buys insurance.

Suppose we now introduce the possibility of insurance, sold with no load. The actuarially fair premium is equal to the expected loss, which is just $0.25 \times 200{,}000 = \$50{,}000$. Thus, anyone who buys insurance pays a premium of $50,000 and has guaranteed wealth of (200,000 - 50,000 =) $150,000.[21] The utility of $150,000 held with certainty is just $ln(150{,}000) = 11.92$. Since this is larger than the *expected* utility of doing without insurance, everyone will want to purchase full coverage, and village aggregate utility is thus 1,192, which is higher than before.

[21] If the earthquake does *not* occur, the premium is paid but there are no losses, so wealth is 200,000 - 50,000 = $150,000. If the earthquake *does* occur, the homeowner pays a premium of 50,000, loses 200,000, and then receives a check for the full amount of the loss, again leaving her with $150,000 net.

Table 1: Insurance vs Non-Insurance,
No Individuation (homogenous risk)

Assumptions	
Population Size	100
Individual Wealth, W	200,000
Size of Loss[i]	200,000
Probability of Loss[*]	25%
Utility function, U(W)	*ln*(W)
No Insurance	
Aggregate Expected Loss	5,000,000
Aggregate Expected Utility	858
With Insurance (Pooling)	
Fair Premium	50,000
Wealth, After Premium	150,000
Utility	11.92
Aggregate Utility	1,192

[*]For every individual.

Now imagine that we have access to some technology that generates perfect predictions: instead of each villager facing a 25% chance of having his or her home destroyed, we know with certainty which 25 homes will be destroyed and which 75 will escape any damage. The owners of the 75

known-to-be-safe houses will obviously have no demand for insurance at any positive premium, since they would be paying for coverage that would be of no use to them. Conversely, owners of the 25 known-to-be-destroyed houses will certainly want insurance. But the only coverage available to them will be at the fair premium for a certain-to-be-destroyed house (100%×200,000 =) $200,000, and there is no reason to buy coverage when the premium is equal to the actual loss.[22] So once the forecasting technology is made available, nobody will purchase insurance.

The loss of risk-spreading that accompanies perfect forecasting leaves the community as a whole worse off.[23] Aggregate welfare is now the same as in the no-insurance state described earlier (858), which is 28% lower than when insurance is possible. Before the technology is introduced, behind Rawls' veil of ignorance, the community would want to ban its use. Too much information can reduce welfare.[24]

B. HOW GOOD CAN BIG DATA BE?

---

[22] Note that it is irrelevant whether the insurance company has direct access to this technology or not. Suppose homeowners are the only ones who know whether or not their house will be destroyed; by the logic above, those who want to buy insurance are only the owners who know they will lose their house for sure. The insurance company can thus infer that anyone who demands insurance will be a guaranteed house-loser, and will price its product accordingly. *Cf.* Alexander Tabarrok, *Genetic Testing: An Economic and Contractarian Analysis*, 13 J. HEALTH ECON. 75, 75–76, 79–82 (1994) (providing an example of this concept in the genetic testing context).

[23] In fact, it in some sense destroys the meaning of "community." Before the forecast, everyone in the village was subject to the same risk, and all had a common interest in minimizing its effects via mutual insurance. After the forecast, however, those who will be spared are no longer interested in sharing their fortune with that of their known-to-be-less-fortunate neighbors.

[24] For an elegant discussion of the divergence between the private and social value of information, see Jack Hirshleifer, *The Private and Social Value of Information and the Reward to Inventive Activity*, 61 AMER. ECON. REV. 561 (1971). Hirshleifer's point is that in a pure exchange economy, "*the community as a whole obtains no benefit . . . from either the acquisition or dissemination of private foreknowledge*." *Id.* at 565 (emphasis in original). Foreknowledge is defined as the accurate prediction of an event that will eventually come to pass (or not), as distinguished from the discovery of something new that need not be discovered at all. *See, e.g.*, *id.* at 562. In my example, information prevents risk-spreading, and hence is actually destructive of social welfare.

Speaking very broadly, Big Data can generate better predictions by uncovering new independent variables, or combinations of variables, that help explain the outcome of interest, and it can help uncover new ways in which the independent variables are related to the outcome.[25] But for most risks for which people seek insurance, it seems virtually impossible that any feasible improvement in the technology of prediction could so significantly increase accuracy as to make insurance impossible.

Assertions of seemingly miraculous predictions emerging from Big Data are often, on closer examination, grossly exaggerated. Two years ago, for example, *New York Times* reporter Charles Duhigg wrote a widely-discussed article about how Target was able to use Big Data techniques to predict, on the basis of their purchasing patterns, which customers were pregnant.[26] The story featured an account of an angry father whose teenage daughter received ads for diapers and wipes, even though (he believed) she was not pregnant. But it turned out that she actually *was*, and Target had apparently used Big Data to figure this out before he did.

Writing in the *Financial Times*, economist Tim Harford effectively debunks this story, however. It turns out that the reported success of Target's algorithm ignored the false positive problem: we didn't get to hear the stories about women who received coupons for babywear but who *were not* pregnant.

> Hearing the anecdote, it's easy to assume that Target's algorithms are infallible–that everybody receiving coupons for onesies and wet wipes is pregnant. This is vanishingly unlikely. Indeed, it could be that pregnant women receive such offers merely because everybody on Target's mailing list receives such offers. We should not buy the idea that Target employs mind-readers before considering how many misses attend each hit.[27]

---

[25] For a brief and appropriately skeptical view of the strengths and weaknesses of Big Data, see Sendhil Mullainathan, *Why Computers Won't be Replacing You Just Yet: A 25-Question Twitter Quiz to Predict Retweets,* N.Y. TIMES (July 1, 2014), http://www.nytimes.com/2014/07/03/upshot/a-25-question-twitter-quiz-to-predict-retweets.html.

[26] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[27] Harford attributes this insight to statistician Kaiser Fung. Tim Harford, *Big Data: Are We Making a Big Mistake?*, FIN. TIMES (Mar. 28, 2014), http://on.ft.com/P0PVBF.

C. IS THE TMI PROBLEM A REALISTIC CONSEQUENCE OF BIG DATA?

It is possible that Big Data may produce too much information, leading to the selective destruction of insurance markets. But is this theoretical possibility one we should be worried about? Although there may be some exceptions, I think the answer for most risks we care about is "no."[28]

For an example of how difficult prediction can be, consider forecasting someone's future earnings at the time they graduate from high school. Economists Alan Kreuger and William Bowen attempted this exercise, considering "an embarrassingly long list of [108] explanatory variables . . . including sets of variables measuring family income, parents' education, parents' occupation, students' expected occupation [on graduating from high school], race, sex, religion, age, and achievement test scores."[29] "Perhaps surprisingly," the authors conclude, "an ordinary least squares regression with these variables accounted for only one-quarter of the variability in earnings."[30] Big Data techniques could be used to reduce the list of 108 variables to a smaller number that were the most powerful explanatory factors. They could be used to find additional variables that might enable some further gains in predictive accuracy. But they cannot dramatically improve the prediction of events or outcomes with millions of independent causes, each of which contributes only a tiny share of the overall effect.

Suppose instead that we are trying to explain whether individual *i*'s house burns down over some fixed period. We might start with traditional underwriting information: the date the house was built, the kind of materials used, the owner's smoking status, and so on. Now consider expanding the set of possible explanatory variables, augmenting traditional underwriting data with new information of the kind Big Data techniques are designed to discover and utilize, such as the homeowner's high school GPA; the list of magazines she subscribes to; and the number of calls made

---

[28] Kenneth S. Abraham & Pierre-André Chiappori, *Classification Risk and Its Regulation*, *in* RESEARCH HANDBOOK ON THE LAW AND ECONOMICS OF INSURANCE (Daniel Schwarcz & Peter Siegelman eds., forthcoming 2015).

[29] Alan B. Krueger & William G. Bowen, *Policy Watch: Income-Contingent College Loans*, 7 J. ECON. PERSP. 193, 196 (1993).

[30] *Id.*

from the home to area code 510.

It is possible that one or more of these new variables, separately or interacted with each other or existing variables, could improve predictive accuracy. For example, when it comes to predicting the chance of a fire this year, knowing that the homeowner had GPA of 2.3 or that she subscribes to *Soldier of Fortune* might be more useful than knowing that her home was built in 1956.

Big Data methods allow the researcher to consider many more variables and combinations of variables than has traditionally been possible, including "high dimensional" cases where the number of explanatory variables is greater than the number of observations.[31] When analysts are searching for a parsimonious group of a few explanatory variables from among many possibilities, Big Data and machine learning techniques can be extremely useful. But that is *not* the same as saying that Big Data can explain the otherwise inexplicable.

There is no doubt that there may be gains to be achieved from using Big Data techniques to predict fire risk. But as Table 2 makes clear, it is almost algebraically impossible that any newly discovered variable (e.g., homeowner's GPA) or combination of variables (Female & Subscribes to *Soldier of Fortune* magazine & GPA less than 2.5) could enable highly-accurate predictions of fire risk. Imagine that, by using Big Data, we found that being female, subscribing to *Soldier of Fortune*, and having a high school GPA of less than 2.5 are collectively associated with a 100-fold increase in fire risk. If Big Data techniques could generate a robust improvement in prediction of this magnitude, it would be truly

---

[31] For an introduction to the theory and some examples, see Alexandre Belloni et al., *High-Dimensional Methods and Inference on Structural and Treatment Effects*, 28 J. ECON. PERSP. 29, 33-34, 38-41 (2014). Moreover, these techniques are designed to prevent "over-fitting" or ad hoc specifications in which the researcher develops an explanatory model that fits the data for a given sample, but is useless for predictive purposes outside of the sample. Overfitting of this kind is more likely as the ratio of explanatory variables to observations increases. In the limit, there are exactly as many variables (plus a constant) as there are observations. In this case, the ordinary least squares estimator will fit the data perfectly, returning an R^2 of one. However, using the estimated model is likely to result in very poor forecasting properties out-of-sample because the model estimated by least squares is overfit: the least-squares fit captures not only the signal about how predictor variables may be used to forecast the outcome, but also fits the noise that is present in the given sample, and is not useful for forming out-of-sample predictions. *Id*. at 30.

shocking.[32] But even if such an improvement were achievable, it would only raise the probability of a fire (for the small number of persons in this group) from 9/10,000 to 900/10,000, which is still less than 10 percent. A dramatic increase from the baseline case, to be sure, but nothing remotely approaching a risk so high as to be virtually certain, one that would shred the veil of ignorance needed to make risk-spreading possible.

Table 2: Back-of-the-Envelope U.S. Fire Risk	
236,200	annual average one- and two-family residential fires in the period 2009-2011.[33]
90,742,000	single unit homes.[34]
0.0026	annual probability that a house will experience a fire (26/10,000)

But what about rare medical conditions, such as Huntington's disease, you might ask? Estimates apparently vary quite widely, but one recent study estimated the annual incidence of Huntington's disease to be 0.38 per 100,000, which is only 1/685th as high as the US annual house-fire risk.[35] Yet some scholars have suggested that Huntington's is essentially uninsurable[36] because it is almost perfectly predictable based on

[32] By "robust," I mean that the correlation would hold up in the future, and would reflect not just a random association in the sample of cases on which the predictive model was estimated. In Mullainathan's example, a Big Data algorithm predicted "which [of a given pair of] tweet[s] gets retweeted [more often] about 67 percent of the time, beating humans, who on average get it right only 61 percent of the time." Mullainathan, *supra* note 25. Impressive as this is, it represents only a 10% improvement (6%/61%) over human performance.

[33] Nat'l Fire Data Center, U.S. Fire Admin., *One- and Two-family Residential Building Fires (2009-2011)*, 14 TOPICAL FIRE REP. SERIES 1, 1 (Sept. 2013), *available at* http://www.usfa.fema.gov/downloads/pdf/statistics/v14i10.pdf.

[34] *Table C-01-AH: General Housing Data—All Housing Units*, H150/11 AM. HOUSING SURV. FOR U.S.: 2011 at 3 (2013), *available at* http://www.census.gov/content/dam/Census/programssurveys/ahs/data/2011/h150-11.pdf.

[35] "Meta-analysis of data from four incidence studies revealed an incidence of 0.38 per 100,000 per year," while a meta-analysis of eleven studies suggested that "[t]he [lifetime] service-based prevalence of HD . . . in Europe, North American [sic], and Australia, . . . [was] 5.70 per 100,000." Tamara Pringsheim et al., *The Incidence and Prevalence of Huntington's Disease: A Systematic Review and Meta-Analysis*, 27 MOVEMENT DISORDERS 1083, 1083 (2012).

[36] Pierre-André Chiappori, *The Welfare Effects of Predictive Medicine*, *in* COMPETITIVE FAILURES IN INSURANCE MARKETS: THEORY AND POLICY

genetic screening: the disease occurs because of a trinucleotide repeat, and anyone with more than 40 repeats is certain to be affected.[37]

For insurance purposes, the relevant difference between Huntington's risk and house fire risk is not their relative magnitudes. Rather, it is that Huntington's has a single, identifiable predictor, the genetic defect is the only source of the condition, and everyone with the defect develops the disease. House fires, by contrast, are not mechanically linked to any single predictable-in-advance cause. Many women have low high school GPAs and read *Soldier of Fortune*, but even in our hypothetical world, only a small fraction of them will experience a house fire. The social world is inherently more complex than the bio-physical world in this respect. And even many medical conditions are more like type-2 diabetes than like Huntington's disease: they are the result of a complicated and poorly-understood mix of environmental and biological factors, and there is simply no clear-cut causal structure that explains when the risk will materialize and when it won't.[38]

The bottom line is that Big Data techniques are not all that useful for single-predictor risks such as Huntington's disease, the cause of which was discovered using ordinary scientific methods. And however useful they are for more complex predictive structures, Big Data techniques do not permit accurate prediction of multiply-caused rare events. While even

---

IMPLICATIONS 55, 56, 65–66 (Pierre-André Chiappori & Christian Gollier eds., 2006).

[37] The defect involves the repetition of a group of three nucleotides (CAG: Cytosine, Adenine and Guanine). Healthy people have between 7 and 35 repetitions of this group. However, an incidence of more than 40 repetitions leads to the presence of the disease. Francis O. Walker, *Huntington's Disease*, 369 LANCET 218, 220 (2007). The condition is autosomal dominant, which means that a defective gene inherited from either parent is sufficient to cause the disease. *Id.*

[38] Consider diabetes (which is actually several different conditions). "Most patients with type 2 diabetes [which "accounts for 80% to 90% of cases of diabetes in the United States"] . . . have some degree of tissue insensitivity to insulin attributable to several interrelated factors . . . . These include putative (mostly as yet undefined) genetic factors, which are aggravated in time by further enhancers of insulin resistance such as aging, a sedentary lifestyle, and abdominal visceral obesity. Not all patients with obesity and insulin resistance develop hyperglycemia, however." Umesh Masharani & Michael S. German, *Chapter 17: Pancreatic Hormones and Diabetes Mellitus*, *in* GREENSPAN'S BASIC & CLINICAL ENDOCRINOLOGY (David G. Gardner & Dolores Shoback eds., 9th ed. 2011), *available* at http://accessmedicine.mhmedical.com/book.aspx?bookid=380.

small improvements in predictive accuracy can be quite valuable,[39] it seems highly unlikely that Big Data techniques will produce dramatic improvements in prediction. Mathematician Jordan Ellenberg recently put it this way:

> There are lots of . . . problems where supplying more data improves the accuracy of the result in a fairly predictable way. If you want to predict the course of an asteroid, you need to measure its velocity and its position . . . The more measurements you can make of the asteroid and the more precise those measurements are, the better you're going to do at pinning down its track. But some problems are more like predicting the weather[,] [because weather is, in the technical sense of the word, *chaotic*.] . . . [H]uman behavior [is] even harder to predict than the weather. We have a very good mathematical model for weather, . . . [but] [f]or human action we have no such model and may never have one.[40]

IV. WHAT IF INSURERS KNOW MORE THAN INSUREDS DO ABOUT INDIVIDUAL RISK?

Even if Big Data methods are not sufficient to generate perfect (or even very good) predictions, they could well have other effects that would be worth taking seriously. Since policyholders themselves are not very good at predicting their own riskiness in many situations, Big Data techniques might offer insurers an improvement on the status quo that allows them to out-predict their customers. As we saw earlier, the economic theory of insurance suggests that market equilibria are highly sensitive to small changes in underlying assumptions or parameters, so things might look very different if insurers were able to use Big Data techniques to discover more about policyholders' riskiness than the

---

[39] Netflix offered a $1M prize to anyone who could improve its movie-recommending algorithm by more than 10 percent. According to a Netflix official, a 10% improvement in their recommendations, small as that seems, would recoup the million in less time than it takes to make another *Fast and Furious* movie. JORDAN ELLENBERG, HOW NOT TO BE WRONG: THE POWER OF MATHEMATICAL THINKING 166 (2014).

[40] *Id.* at 164-65.

policyholders themselves knew. Thus, whether or not these methods yield good predictions in some absolute sense, they could still profoundly shape equilibria, even if all they do is improve insurers' predictions *relative* to what insureds know.[41]

What follows is an attempt to illustrate this relatively simple observation.

A. CHARACTERIZING INFORMATION: WHO KNOWS WHAT

Consider a very simple description of possible information stocks. Policyholders face a known loss, L, which is the same for everyone. Each policyholder *j* has a unique probability of experiencing this loss, p_j. The actuarially fair premium for policyholder *j* is equal to *j*'s expected loss:

$$E(L) = p_j \times L.$$

In turn, the probability of loss depends on facts about the policyholder, which we can describe as a vector of characteristics, X_j. We can thus write

$$p_j = f(X_j),$$

which says nothing more than that the probability that individual *j* will experience a loss is a function of the value of the various explanatory variables for that individual, X_j.

We can go further and partition the variables that make up X_j into two possibly-overlapping parts. $X_{j,P}$ represents all the information the policyholder knows about himself—for example, how recklessly he drives. $X_{j,I}$ represents the insurer's information about *j* (for example, the riskiness of *j*'s car, or of the area where he typically drives). Some information will, of course, be uniquely held by one party, while some will be common to both (*j*'s sex or age). In addition, we should allow for information that is known to nobody, which we can denote as random error, ε. Thus, the expected loss (and fair premium) for policyholder *j* can be written as:

$$E(L) = f(X_{j,P}, X_{j,I}, \varepsilon)L.$$

---

[41] Two hikers spot a bear getting ready to charge them. The first hiker drops his pack, takes off his hiking boots, and begins to put on running shoes. The second hiker asks, "What's the point? You're never going to outrun that bear." The first replies: "You're right, I won't; but all I need is to outrun *you*."

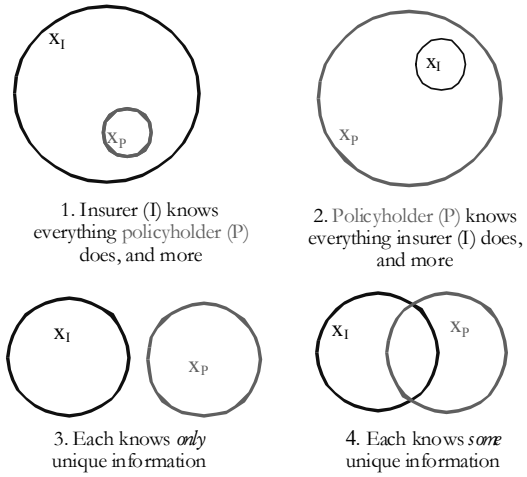

Figure 1

Figure 1 presents some possible configurations of information sets. For example, in panel 1, the insurer knows everything the policyholder knows, as well as some information in addition. In panel 2, the situation is reversed; the policyholder knows everything the insurer knows, and more.

It has generally been assumed by economists that (2) is the best description of how the world works. For example, all models of adverse selection and moral hazard are based on this characterization. While it may seem implausible, there is actually a sophisticated justification for this assumption. When the insurer quotes a price for insurance coverage for individual *j, j*'s premium, it will presumably make an optimal computation of *j*'s riskiness, based on all the information it has at its disposal. So the insurer's estimated fair premium for *j* will be $f(X_{jI})\times L$. But that's just the expected loss for policyholder *j*, given the information available to the insurer, X_{jI}. And since the premium is actuarially fair,[42] policyholder *j* can easily deduce what the insurer thinks his risk of loss must be. For example,

[42] This is required in a competitive equilibrium. A premium that is less than actuarially fair can be expected to earn losses, and the insurer will prefer not to offer any policy at all than to offer one that loses money. A premium priced *above* the actuarially fair level will attract competitors who can offer a slightly lower price and lure away all customers. So the only sustainable price in a competitive market is the fair premium.

suppose the loss is known to be 100. Then a quoted premium of 2 implies that the insurer must believe there is a 2% chance it will have to pay out 100. That, in turn, suggests that even if the policyholder does not know exactly what the insurer knows, he can infer all he *needs* to know about the insurer's information *via* the premium he is quoted, which will necessarily reveal exactly what the insurer believes about the policyholder's expected loss. So the insurer effectively ends up having to surrender all its private information in a competitive equilibrium, while the policyholder doesn't.[43] That situation resembles panel (2) of Figure 1.

But this simple story, appealing as it is, need not be correct. It is possible to have equilibria in which the insurer knows less about insureds than they know about themselves, even with completely rational consumers, a competitive market, zero-cost (no load factor) insurance, and no uncertainty about the size of the loss.[44] The next section explains, by way of an example.

B. EQUILIBRIUM WHEN POLICYHOLDERS ARE BETTER INFORMED THAN INSURERS[45]

Suppose that the population consists of equal numbers of two types of insureds, high-risk and low-risk. The first group has a risk of loss equal to 0.4 (p_H = 40%); the second has a risk of loss equal to 0.3 (p_L = 30%). The loss is known to be 100 for all individuals who experience a loss. The fair premium *for the group as a whole* is just the average loss:

---

[43] The policyholder reveals *some* information when he decides to accept or reject the insurer's offer, but it should be clear that this decision does not give away everything the policyholder knows about his own riskiness.

[44] If consumers are unable to make rational inferences—and the evidence cited suggests this is indeed the case—their ability to extract the insurer's estimate of their own riskiness from the premium quotation they receive is obviously diminished. The ability to extract this information is further diminished by any markup over the fair premium to cover the insurer's cost and by failures of competition to drive prices down to the zero-profit level. KUNREUTHER ET AL., *supra* note 16.

[45] Bertrand Villeneuve, *Competition Between Insurers with Superior Information*, 49 EUR. ECON. REV. 321 (2005), provides the careful analysis on which this loose paraphrase is based. There are important background conditions (e.g., that all policyholders are risk averse enough so that they will demand insurance at each of the possible premiums) which are too technical to consider here.

$$\pi = \left(\frac{1}{2}p_H + \frac{1}{2}p_L\right) \times 100 = 35.$$

Assume further that for any individual *j*, the insurer knows exactly which group *j* is in, while *j* knows only the *average* risk of the population as a whole, but not his own individual risk. The industry contains *N* competitive firms, so that premiums are driven down to the actuarially fair level (given that there are no operating or other costs). Thus, all firms earn zero profit.

Suppose the insurer makes an offer to sell insurance to individual *j* by quoting her a premium.[46] Consider first the possibility that the insurer quotes the group-wide average premium of 35. How would a policyholder react to this offer? If she *knew* she were a low-risk individual, she should reject the offer, because in a competitive market, she would be able to attract a better one from another insurer until the premium was actuarially fair *for a known low-risk individual*. (Conversely, a known high-risk individual would be delighted to be quoted a premium that was less than his actuarially fair value.) But the whole point is that the policyholder does *not* know her own risk type, so the premium of 35 is the best she can expect, given her ignorance of her own riskiness. Thus, both high and low-risk individuals would be content to stick with the average or "pooled" premium, if they were offered it.

But for this to be an equilibrium, we have to establish that the insurer would want to quote the average price in the first place. Consider first what happens when the insurer knows that j is low-risk (but remember, j herself does not). A premium of 35 implies that the insurer would earn profits of 35-30 = 5 for this customer, if she accepts the offer. But if the insurer offers a premium appropriate for the population average risk of 35, it will then be competing with every one of the other N insurers in the market who also offer this price. That in turn means that the insurer faces a 1/N chance of landing this consumer, for an expected profit of 5/N. Alternatively, the insurer might consider quoting a slightly lower premium, say 34, and having a 100% chance of attracting this policyholder given that all its competitors are quoting a price of 35. That would yield a profit of 100%×(34-30) = 4. As long as the number

[46] Significantly, this is what is known as a "signaling" equilibrium because the informed party—here, the insurer—makes the offer. In standard models of insurance market equilibrium, it is the *uninformed* party (still the insurer, but the policyholder knows everything that the insurer does and more, so the insurer is *un*informed) who makes the offer, which leads to a "screening" equilibrium.

of rivals is greater than 2, the insurer would prefer to offer the lower price and land the customer with certainty.

Thus, it might look as if quoting the blended premium (35) cannot be an equilibrium, because an insurer would prefer to do something else. But that intuition turns out to be wrong. Once an uninformed customer receives a quote of 34 from an insurer—who is known to be better informed than she is—she will instantly know that the *insurer* knows she is low-risk.[47] With this knowledge, she is then in a position to demand a reduction in premium to 30 (befitting a known low-risk customer); in a competitive equilibrium with full information on all sides, the zero-profit price is the only one that can prevail.

The point is that by quoting an even slightly more-appropriate price, the insurer ends up telling the consumer exactly what her risk is, and the consumer is then in a position to use that information against the insurer, by insisting on an even lower premium. And in a competitive market, she will, in fact, receive that lower premium. Thus, a small deviation from the blended (average) premium will not be profitable for the insurer. Sticking with the "pooled" rate will be the best the insurer can hope to do.

C. POOLING VS SEPARATION

The careful reader—if he or she has gotten this far—might find something surprising here. A world in which insurers know more about each policyholder than the policyholder does about herself is actually supportive of a *pooling* equilibrium, one in which all consumers pay the same "bundled" or average premium. The *non*-existence of a pooling equilibrium in the presence of adverse selection is one of the key insights of the pioneering Rothschild/Stiglitz model of insurance markets: when consumers know more than insurers do, policyholders' ability to select a policy based on their "inside" information makes a pooling equilibrium unsustainable in a competitive market.[48]

You might think that as insurers learn more and more about their customers, premiums would become more and more individualized and the possibility of pooling would only be diminished. But the weird economics of insurance markets demonstrates that this need not be true. The example above illustrates that when the insurer knows each customer's risk exactly,

---

[47] An offer of 34 is only profitable if made to a known low-risk consumer.

[48] Rothschild & Stiglitz, *supra* note 9, at 639.

while customers know only the group average risk, pooling equilibria *are* possible. Unfortunately, theory predicts that separating equilibria (in which each type pays a premium appropriate to its riskiness) are also possible.[49] So, in the end, the lesson is cautionary. Theory does not support the idea that as insurers learn more about their customers, pricing will necessarily become more individualized and pooling and attendant risk-spreading will necessarily decrease. Instead, a world in which insurers know more about policyholders than the latter know about themselves might actually give rise to *more* pooling.

V. BIG DATA, BIG INSURANCE

In this section, I want to very briefly discuss 2013 Nobel Laureate Robert Shiller's[50] visionary[51] ideas for using Big (or at least More) Data to dramatically increase risk-spreading by allowing consumers to insure (pool) risks that they are currently forced to bear themselves. Shiller's insight is that new kinds of data, aggregated in new ways, could lead to radically new forms of insurance against risks that consumers are currently forced to bear themselves. (This is a somewhat different take on what "Big Data" means, since we are no longer talking about data-mining techniques to extract predictive information from high-dimensional data. Rather, as I explain below, we are concerned with the prospect of creating new *kinds* of information beyond that which is currently available.)

Consider, for example, the risk that one's house might decline in value (something few people did in fact consider in 2003, when Shiller's book was published), or the risk that one's chosen line of work might

---

[49] Villeneuve, *supra* note 45. The existence of separating equilibria depends on the degree of consumers' risk aversion and the difference in riskiness between the two types. Note that despite its complexity, the model admits only an extremely limited degree of consumer heterogeneity. Policyholders differ *only* in their riskiness and not, for example, in their degree of risk aversion. Nor are consumers subject to any behavioral "flaws" or deviations from rationality. For an attempt to incorporate such heterogeneity into a theoretical (simulation) model of insurance markets, see Tsvetanka Karagyozova & Peter Siegelman, *Can Propitious Selection Stabilize Insurance Markets?*, 35 J. INS. ISSUES 121 (2012).

[50] ROBERT J. SHILLER, THE NEW FINANCIAL ORDER: RISK IN THE 21ST CENTURY (2003).

[51] Some have almost gone so far as to suggest that "hallucinatory" would be a better description. *See* Stephen A. Ross, *Review of* The New Financial Order *by Shiller*. 42 J. ECON. LIT. 1098 (2004).

experience a drop in demand, causing a fall in one's earnings. Risk-averse individuals should want these products, which protect against important risks that they would prefer not to fact.

But individualized insurance against these risks cannot work, Shiller points out, because of Moral Hazard.[52] If the value of my home is fully insured, I have an incentive to under-maintain it: maintenance is costly, after all, and my home value insurance policy will cover any drop in price when it comes time to sell the property.[53] Similarly, if my livelihood (earnings) is fully insured, I may slough off because hard work is costly and my livelihood insurance will pick up any shortfall in my paycheck that results from my shirking.[54]

Shiller's brilliant insight is that even if some component of these risks is uninsurable at an individual level, it is possible to create a viable insurance product that covers aggregate-level risks without any moral hazard risks. Thus, instead of insuring against a fall in the value of *my* house, I would buy coverage against a drop in the value of *all* houses in my city or neighborhood. Instead of insuring against a fall in *my own* earnings, I would buy coverage against a drop in the earnings of all persons in my profession (law professor) or perhaps some narrower aggregate (all law and economics professors).

Under Shiller's solution, some risks remains with the consumer, as they must to preserve incentives, but at least medium- to large-scale risks can be insured against. If the largest employer in town closes its factory and all local house prices plummet, I am covered. If nobody wants to go to law school any more, and law professor salaries plunge, I am covered there as well.

The genius of this approach is that it offers maximal insurance with no potential for Moral Hazard, since insurance is offered only against drops in an aggregate (price index), over which no individual exerts any control. If I under-maintain my house, I bear 100% of the marginal loss in value, relative to the average house in my neighborhood. If I slack off rather than working hard, I do less well than the average law professor (even if all salaries drop), and

---

[52] And possibly Adverse Selection as well, although Shiller scarcely mentions adverse selection in his book.

[53] Of course, if it were possible to write an insurance contract that covered exactly what kind of maintenance I was required to do, this problem could be solved. But, it seems clear that maintenance is simply too complicated, heterogeneous and subjective to be captured by an ex ante contract.

[54] *See*, *e.g.*, Soviet-era Russia. There are possible selection issues as well if homeowners know better than their insurers whether their house needs repairs or what their own future work plans entail.

those losses are not covered by my insurer. Shillerian insurance thus preserves maximal incentives for me to work hard and to maintain my home, while permitting me to pool risks that I would like to avoid.55

But in order for this kind of insurance to work, we need "Big" data on aggregates (neighborhood home values, earnings by occupation or sub-specialty). This information would need to be built up from detailed data collected at an individual level. For each house, we have to know its age, its square footage, its condition, and of course its price. This data could then be aggregated to provide quality-weighted neighborhood-level information that could then be used to set premiums and payouts. Shiller and his collaborator Karl Case actually created such a dataset, which is now maintained (for several cities) by the rating agency Standard and Poors.[56]

VI. CONCLUSION

Equilibrium in insurance markets is highly sensitive to seemingly-innocuous details about how offers are made and received, by whom, and under what conditions. Robust predictions about how markets will respond to any exogenous change are very difficult. It would therefore be silly to claim, at least as a theoretical matter, that Big Data will have little or no effect on insurance market equilibria. But at least the notion that Big Data techniques will enable some sort of perfect prediction seems pretty far-fetched.

And while the collection and analysis of additional information may pose some significant privacy concerns, it may also make possible the creation of new markets for spreading risks that rational individuals should greet with approval.

---

[55] There is a structural similarity between this kind of insurance and Robert Cooter's theory of the law and economics of "precaution." *See* Robert Cooter, *Unity in Tort, Contract, and Property: the Model of Precaution*, 73 CAL. L. REV. 1 (1985). In both models, one party (the insurer or the injurer) bears responsibility for the inframarginal precautions, while the other party (the insured or the tort victim) bears responsibility for the marginal precautions, thereby providing simultaneous incentives for both parties to take efficient levels of care.

[56] *See S&P/Case-Shiller Home Price Indices*, S&P DOW JONES, http://us.spindices.com/index-family/real-estate/sp-case-shiller (last visited Aug. 11, 2014).