



Data-enriched prediction of insurance risk

Ruo Jia^a, Shaoran Li^{a,*}, Ye Yin^b

^a School of Economics, Peking University, China

^b School of Insurance, University of International Business and Economics, China

ARTICLE INFO

JEL classification:

C52

G52

Keywords:

Insurance underwriting

Predictor selection

Risk prediction

Big data

ABSTRACT

In this study, we examine the impact of big data and corresponding prediction techniques on insurance risk assessment. We demonstrate that both big data and the Least Absolute Shrinkage and Selection Operator (LASSO) approach, a modern predictor selection technique, effectively improve insurance risk assessment accuracy. In an average of proxies, the out-of-sample health risk prediction accuracy improves by 106 %, in which big data, in addition to traditional insurance policy and demographic information, contributes 82.5 %, and the LASSO model contributes 17.5 %. We also demonstrate that big data obtained from smartphone use offers extra-predictive power in addition to past medical histories. We employ Adaptive Group LASSO to determine that the most fruitful data collection sources for health insurance underwriting include personal digital devices, recent travel experience, and insureds' credit records.

Introduction

Insurance markets inherently grapple with information asymmetry, in which individuals possess more knowledge about their own risk profiles than insurers. This information imbalance leads to adverse selection, as higher-risk individuals are more likely to seek insurance at premiums based on average risk (Akerlof, 1970; Rothschild & Stiglitz, 1976). To mitigate this issue, insurance companies use consumer information, such as age and gender, for risk assessment and underwriting (Cather, 2018), which benefit low-risk consumers as well (Browne & Kamiya, 2012). However, key health predictors other than age and gender remain inaccessible to insurers. Consequently, proxy variables that can capture underlying health conditions effectively are needed urgently.

To address these gaps, insurance companies increasingly have been adopting alternative data for risk assessment, but one critical challenge persists: The sheer volume of alternative data leaves insurance companies uncertain where to invest. Without a systematic approach, insurance companies risk investing in data that provide little predictive power while overlooking meaningful indicators. Our study addresses both challenges urgently. We develop a method to identify useful proxies for key, yet-often-unobservable, health risk predictors using passively collected data. This approach not only bypasses self-reported data's unreliability, but also, through the application of the Adaptive Group LASSO (Least Absolute Shrinkage and Selection Operator) method, allows insurance companies to prioritize the most impactful and cost-effective data streams for collection and analysis.

In this paper, we examine a proprietary critical illness (CI) insurance dataset to examine these questions and demonstrate that big data, coupled with the LASSO-style prediction model, massively enhances health risks' prediction accuracy. We employ two variables as health risk proxies: i) *CI_Claim*, a binary variable equal to one for insureds who filed CI claim(s) during our observation period; and

* Corresponding author.

E-mail addresses: ruo.jia@pku.edu.cn (R. Jia), lishaoran@pku.edu.cn (S. Li), 1258198287@qq.com (Y. Yin).

<https://doi.org/10.1016/j.risk.2025.100028>

Received 9 October 2024; Received in revised form 10 June 2025; Accepted 26 July 2025

Available online xxxx

2950-6298/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

ii) *Medical Claim Frequency*, the number of public health insurance claims filed during the two years preceding the CI insurance application. The combination of big data and the LASSO model yields a notable 106 % improvement in health risk prediction accuracy, as measured based on average improvements calculated from two proxies. Specifically, big data contributes 82.5 % enhancement, while the LASSO model improves accuracy by 17.5 %.

Subsequently, we employ Adaptive Group LASSO to identify informative data collection directions for future health risk assessment and health insurance underwriting. We find that—beyond demographic and insurance policy information—personal digital devices, recent travel experience, and insureds' credit records are the most fruitful categories of information in predicting health risks, with an emphasis on the importance of both financial and nonfinancial dimensions of insureds to predict health risk and health insurance claims.

This paper's contributions are threefold. First, we contribute new predictors, namely smartphone-based information, for insurance risk assessment. Extant literature on underwriting/pricing focuses on identifying new predictors. Black et al. (1994) document that academic performance, occupation, financial trading activities, and income level are useful for risk assessment in life and health insurance. Finkelstein and Poterba (2004) find that in the U.K. annuity market, location information predicts insureds' longevity risk. Fong (2015) find that—beyond age and gender—education, marital status, and BMI also are correlated to insureds' longevity risk in the U.S. annuity market. Geyer et al. (2020) show that driving distance, number of car rides, and speeding affect auto insurance risk. In the era of big data, obtaining the above information is less of a problem than properly weighing their importance in a prediction model. In this paper, we apply the (Group) LASSO model to select the most efficient predictors for the prediction model, optimizing in-sample and out-of-sample performance with explicit interpretations. We use real-world, smartphone-based granular data to demonstrate the extra-predictive power of big data (also called alternative data). Based on real claim experience, we justify and quantify the value of insurers' efforts to collect insureds' digital footprint and smartphone use behaviors as new risk predictors (Guillen et al., 2021). We also focus on the value of additional big data that has not been fully exploited yet in extant literature.

Second, our study contributes to the risk prediction methodology. In the big data era, avoiding “overfitting” by properly using this abundant information and weighing its importance in a prediction model are critical. For example, Powers and Xu (2025) stress the importance of avoiding “idiosyncratic overfitting,” which arises when a chosen parametric distribution's functional form is overly complex for the data. A large body of extant studies has stressed the extra-predictive power elicited by machine learning methods. For example, Fuster et al. (2021) argue that the machine learning method triangulates excluded characteristics. Boudt et al. (2025) presents a novel, data-driven, news-monitoring system designed to detect risks for AML supervision of financial institutions. This paper is conceptually close to Qu et al. (2025), who examine China's new Internet auto insurance market and demonstrated that applying machine learning models (e.g., k-nearest neighbor and gradient boosting decision tree) with telematics and behavioral data to classify policyholders into more homogeneous risk groups could reduce adverse selection. Our paper focus on China's CI insurance market and aim to apply the LASSO-style model to select the most fruitful data collection directions.

Third, we guide the direction of the most fruitful data collection for health risk prediction, considering that data are expensive. It remains unclear how to select potential predictive alternative data systematically and to what extent and which segments of these alternative data are most useful. Adaptive Group LASSO lets data speak by grouping them based on their characteristics and collection sources, and then selecting the most relevant groups, thereby indicating the most valuable areas to dig for additional underwriting information.

This paper is organized as follows. In Section 2, we present the (Adaptive Group) LASSO prediction models. In Section 3, we introduce big data for insurance risk assessment. In Section 4, we report out-of-sample prediction results. In Section 5, we provide robustness checks. Finally, in Section 6, we conclude the paper.

Prediction models

In the big data era, variable selection is as crucial as variable collection. We extend predictive modeling by leveraging LASSO and Adaptive Group LASSO to refine both individual variable selection and category-level grouping. LASSO ensures a sparse model that mitigates overfitting in out-of-sample predictions, enhancing both interpretability and robustness while its initial estimates guide Adaptive Group LASSO in consistently identifying the most informative data exploration areas (Wang & Leng, 2008).

LASSO

The LASSO-style variable selection method is the most popular way to select relevant variables in a high-dimensional context (e.g., Tibshirani, 1996). Suppose that we observe one set of sample data (Y, P) , in which Y presents the $n \times 1$ vector of dependent variables, while P denotes the $n \times P$ matrix of potential covariates $(X_1, X_2, \dots, X_P)$, allowing for $P > n$. Tibshirani (1996) propose this method to conduct both variable selection and regularization in the linear model in high-dimensional cases.

$$\min_{\alpha} \sum_{i=1}^n \left(Y_i - \sum_{p=1}^P \alpha_p X_{ip} \right)^2 + \lambda_n \sum_{p=1}^P |\alpha_p|, \quad (1)$$

In Eq. (1), Y_i and X_{ip} are the i th element of Y and X_p , respectively. λ_n is a data-driving tuning parameter, and the attractive property of LASSO is that it can achieve initial selection by shrinking some $\alpha = 0$ and estimation even if $P > n$.

However, the estimates in Eq. (1) are shrunk and, therefore, biased. Thus, Belloni and Chernozhukov (2013) suggest using ordinary least squares (OLS) and employing only LASSO-selected regressors to make out-of-sample predictions to avoid the “overfitting” problem.

We select Q variables out of P as the most powerful predictors to achieve two goals: (1) We avoid the “overfitting” problem caused by high-dimension P because many irrelevant predictors may provide high in-sample R^2 , but less-efficient, out-of-sample predictions

(2). Compared with other machine learning and deep learning methods, the LASSO method can present the prediction power source explicitly, i.e., selected predictors or predictor groups rather than a “black box” with ambiguous model specifications.

Adaptive group LASSO

An ideal prediction model should not only select the most predictive variables, but also identify the most fruitful areas for data collection. This is particularly true considering that insurance underwriting data collection, which includes data processing, is very expensive. For the latter purpose, we introduced “group LASSO” (Yuan & Lin, 2006). Suppose that there are L latent groups in $(X_1, X_2, \dots, X_P)$, and variables can be grouped based on their characteristics or collection source. The variables then are selected in a group manner. Namely, for variables in the same group, group LASSO either keeps or discards them all. The selection results by group LASSO indicate which kinds of variables are valuable and important to predict the dependent variable and, therefore, justify more resources to be allocated to this area for better risk prediction.

Without loss of generality, we assume that there are H variables in each group, with L groups altogether, then the selection model is written as:

$$L_n(\mu, \beta_n) = \sum_{i=1}^n \left[Y_i - \mu - \sum_{l=1}^L \sum_{h=1}^H \beta_{lh} X_{ih} \right]^2 + \lambda_n \sum_{l=1}^L \hat{\omega}_l \|\beta_l\|_2, \quad (2)$$

in which λ_n is a tuning parameter, while $\|\beta_l\|_2$ is the L^2 norm of the l^{th} coefficient vector, $\beta_l = (\beta_{l1}, \dots, \beta_{lH})^T$, and

$$\hat{\omega}_l = \begin{cases} \|\tilde{\beta}_l\|_2^{-1} & \text{for } \|\tilde{\beta}_l\|_2 > 0 \\ \infty & \text{for } \|\tilde{\beta}_l\|_2 = 0 \end{cases},$$

in which $\tilde{\beta}_l$ is an initial and consistent estimator. The non-zero estimate $\tilde{\beta}_l$ with $\|\tilde{\beta}_l\|_2 > 0$. This idea is called Adaptive Group LASSO (Wang & Leng, 2008) because it utilizes estimates from the preliminary estimation as penalty parameters.

There are two steps to select the most relevant K groups out of a large set of candidate variables, P , for predicting the risk proxy, Y : (1) Substitute all P candidate variables into LASSO selection. However, one drawback of LASSO is that it may leave plenty of small, but non-zero, coefficients, which blur the important variable categories. Nevertheless, the first step provides crucial hints that are helpful to the discriminatory penalty in the next step. (2) Use *Step 1* estimates to construct penalty-weighting coefficients for all L groups and substitute all groups into Adaptive Group LASSO.¹ *Step 2*'s results provide a clear conclusion about which categories of variables are most relevant and valuable to predict Y .

A coin has two sides, and so does Adaptive Group LASSO, as it may omit important predictors from other less-important and, thus, unselected groups, while overemphasizing predictors from selected groups. Therefore, we use variables selected from *Step 1* for prediction purposes and the results from *Step 2* to identify the most fruitful information collection directions.

LASSO-style model advantages

We use LASSO-style models in this paper over other regularization models, such as ridge regression or elastic net, for two reasons. First, this paper prioritizes model interpretability and explicit variable selection, and LASSO's ability to enforce sparsity by setting irrelevant coefficients to zero is critical for this objective. This is a distinct advantage over ridge regression, which does not conduct variable selection in this manner. Second, LASSO is uniquely suited for this paper's two-step methodology, as theoretical evidence supports how initial LASSO estimates effectively can guide subsequent Adaptive Group LASSO in consistently identifying the most informative data exploration areas. Therefore, LASSO's ability to provide both a sparse model and a theoretically sound foundation for the group-level analysis makes it the most appropriate choice for the paper.

Big data

Datasets and institutional background

We obtain our dataset from InsurTech, a Chinese company. InsurTech offers underwriting suggestions to insurance companies in an online private insurance purchase environment: An insurance applicant selects a private insurance plan from an insurance company, then the applicant checks the health declaration, fills out personal information, and receives a premium quotation. To confirm the application, the applicant needs to authorize the use of personal information, including those provided and those to be collected additionally for underwriting purposes. Right after the applicant confirms the application and before the payment option pops up, InsurTech receives underwriting information filled out by the applicant, obtains additional alternative data in real time with the applicant's authorization,² and uses its proprietary model to generate underwriting suggestions to the insurance company. The

¹ This is why we term it Adaptive Group LASSO, as it uses the preliminary information to weight the penalty, which leads to better selection results.

² In China, customers can authorize insurance companies and their third-party partners to access necessary information for underwriting. This authorization facilitates collection of relevant data while ensuring compliance.

applicant then could pay the premiums to complete the application, and InsurTech receives the underwriting and validation dates from the insurance company if it accepts the application.

In our context, although applicants provide consent to the insurance company and its third-party partners (e.g., InsurTech) for data collection, they do not know which exact data are collected for underwriting and pricing. This minimizes the possibility that an insured changes their daily digital footprints for purposes of a better underwriting outcome.

The dataset comprises CI insurance applications, and the policy application date is between January 1, 2019, and January 31, 2021. The dataset comprises three sources: policy-level data (i.e., traditional data); smartphone-based label data additionally obtained by InsurTech (i.e., alternative data, big data); and claim data from public health insurance.³

The policy-level data contains demographic information on (potential) insureds and insurance policy details, such as insurance amount, policy duration, premiums, and the insured provinces, among other data. Our observation window for CI claims ended in March 2022. The *CI_Claim* equaled 1 if the insured files a claim by March 2022 and 0 otherwise. We consolidate application-level observations to insured-level observations.

The smartphone-based label data contain alternative data that InsurTech collects from data vendors about potential insureds, including additional demographic information (e.g., occupation), digital device information (e.g., device brand), location, messages, and notifications containing specific keywords and types of apps installed, as well as personal credit history inquiries from loan institutions.

The claim data from public health insurance contains two-year claims preceding a CI insurance application, as well as claim records from secondary and tertiary hospitals of potential insureds in 13 provinces: Inner Mongolia; Beijing; Jilin; Sichuan; Tianjin; Shandong; Guangxi; Jiangsu; Henan; Hubei; Hunan; Fujian; and Chongqing. In Online Appendix Table A1, we present the GDP rank and life expectancy in each province in 2020 to demonstrate that these provinces present no significant selection bias regarding economic and health conditions. Considering the COVID-19 pandemic's potential influence on our results, we excluded Hubei province in a robustness check (see Table 6).

Variables

We employ two variables to proxy potential insureds' health risk: (1) *CI_Claim*, a binary variable equal to 1 for insureds who claimed a CI by March 2022 (concerning different claim observation windows of different insurance policies, we set a one-year observation window for all policies in a robustness check; see Table 6); and (2) *Medical_Claim_Frequency*, the number of claims from public health insurance during the two years preceding CI insurance applications.

We classify 63 predictors into 19 groups based on data collection sources and economic rationale. Definitions and descriptions of variables are presented in Online Appendix Table A2. We classify predictors from the policy dataset into three groups. (1) *Demographics*: age and gender of the insured; (2) *Insurance plan*: maximum insurance amount applied, additional coverage of the premium waiver if the applicant is sick, whole life coverage endorsement, the number of different insurance products applied, and total number of applications; and (3) *Applicant-related*: age and gender.

We classify predictors from the label dataset into 16 groups based on their characteristics and data collection sources:

- (4) *Additional demographics*: whether to have a child.
- (5) *Occupation*.
- (6) *Digital device used*: device brand; telecom company; and device type.
- (7) *Borrowing history*: an indicator of whether the insured received any notification that they borrowed money and the number of small loan apps installed.
- (8) *Investment*: an indicator of whether the insured received any notification that they bought financial/wealth management products and the number of money management apps installed.
- (9) *Travel experience*: an indicator of whether the insured received any notification indicating that they drive, and the number of provinces/cities that the insured visited in 180 days before the application.
- (10) *Occupation inferred from messages and apps*: functional apps installed and indicators of whether the insured received specific notifications indicating their specific occupations.
- (11) *Living characteristics*: an indicator of whether the insured use apps in midnight, the number of fitness-related apps installed, and an indicator of whether the insured received any notification that they bought a lottery.
- (12) *Parenthood*: the number of maternal apps installed and the number of children education apps installed.
- (13) *Insurance purchase*: an indicator of whether the insured received any notification that they purchased insurance.
- (14) *Number of financial institutions' inquiries*: numbers of financial institutions who checked the potential insured's credit history in the last 30, 60, 90, 180, and 240 days before application.
- (15) *P2P institutions' inquiries*.
- (16) *Installment loan inquiries*.
- (17) *Payday loan inquiries*.

³ Policy data and label data are at the application level, and the public health insurance claim dataset is at the claim level. We construct the individual-level public health insurance claim variables and then link them with policy data and label data using the insured's identifier to form the combined dataset.

(18) *Other financial institutions inquiries*: Finally, we classify the number of times that financial institutions checked an insured's credit history in the last 30, 60, 90, 180, and 240 days.

(19) *Financial institutions' check time*.

In our prediction model, we also included fixed effects of potential insureds' provinces, as well as application months and years.

Sample construction and summary statistics

We constructed two samples for the two health risk proxies. Sample A is for *CI_Claim*: We keep potential insureds aged 18–60 and those whose alternative data and claim status are available. Sample A contains 55,560 observations. Sample B is for *Medical_Claim_Frequency*: In Sample A, we include individuals whose insured provinces are among the 13 provinces with claim data from public health insurance available. This additional step yields Sample B's 35,967 observations.

Table 1 summarizes both samples. *CI_Claim* was a rare sample, in which claimants accounts for 0.28 % of the insureds. The average *Medical_Claim_Frequency* is 1.5. The average insured age for both samples is around 33, and about 45 % of insureds are male. The alternative data distributions are similar across the two samples.

Notably, our dataset have several limitations. First, the 13 provinces within China from which our data originates spanned the years 2019–2021. While this dataset provides a valuable window into insurance risk prediction in a significant portion of the Chinese market, it may not be fully representative of all provinces nationwide due to potential regional variations in healthcare access, lifestyle factors, and economic conditions. Furthermore, a large part of our data collection time frame overlapped with the COVID-19 pandemic. Although we exclude Hubei province, the epicenter of China's outbreak, in the robustness check, the pandemic's impact was felt nationwide to varying degrees during this period, potentially influencing health behaviors, healthcare utilization, and consequently insurance claim patterns within our dataset. These limitations should be considered when interpreting our findings and generalizing them to broader contexts.

Empirical analyses

Out-of-sample prediction

In this section, we discuss how we conduct out-of-sample predictions to assess the extra-predictive power of alternative data and the LASSO approach. We randomly separate our dataset into the training sample (30 % observations) and testing sample (70 % observations).⁴ We estimate the model in Eq. (3) using the training sample and use the estimated coefficients to predict y_i for the observations in the testing sample.

$$\hat{y}_i = \hat{\beta}_1 \times \text{Traditional_data}_i + \hat{\beta}_2 \times \text{Alternative_data}_i, \quad (3)$$

where y_i refers to *CI_Claim*_{*i*} or $\ln(1 + \text{Medical_Claim_Frequency}_i)$ of individual *i*. *Traditional_data*_{*i*} refers to *Age*, *Age*², *Age*³, gender, and the interaction of age and gender of individual *i* to account for the nonlinear relationship between health risk and traditional data (e.g., Shi et al., 2018). *Alternative_data*_{*i*} refers to the potential risk predictors attained from the big data in Groups (2) to (19), as described in Table 1. For categorical variables, we chose the most popular categories as the baseline and omitted them. We use the Area Under the Receiver Operating Characteristic Curve (AUC) as the prediction accuracy measure when *CI_Claim*, a binary variable, is the dependent variable, but use the out-of-sample *R*² when *Medical_claim_frequency* is the dependent variable. We repeat the above procedure 100 times. The mean AUC and out-of-sample *R*² are present in Table 2.

To decompose contributions from alternative data and the LASSO model, we compared the following three models' performance: (1) OLS using traditional data (benchmark)⁵; (2) OLS using traditional and alternative data (to quantify the contribution from big data); and (3) LASSO using traditional and alternative data (to quantify the LASSO contribution).

As present in Table 2, based on an average of the two-risk proxy specifications, the combined application of big data and the LASSO model enhance health risk prediction by 106 %. Notably, big data contributes an 82.5 % improvement in predictions, emphasizing its substantial extra-predictive power. The LASSO model further refines predictions, contributing an additional 17.5 % improvement. This decomposition demonstrates that big data increases predictability significantly, while the LASSO model, an appropriate modeling technique for big data, provides extra-predictive power.

In Column (1) of Table 2, using *CI_Claim* to proxy health risk, the baseline out-of-sample prediction with OLS and traditional data yields an average AUC of 0.622 (the average of 100 repetitions, similarly hereinafter). The best out-of-sample prediction utilizes linear regression with LASSO and both traditional and alternative data, resulting in an average AUC of 0.696, a 61 % improvement in prediction accuracy compared with the benchmark prediction (i.e., $[0.696 - 0.622] / [0.622 - 0.5] = 0.61$). OLS regression with traditional and alternative data achieves an average AUC of 0.673. Big data's contribution is 69 %, while the LASSO model contributes 31 %. In Column (2), using *Medical_Claim_Frequency*, the baseline model remains the same, with an average out-of-sample *R*² of 0.0196. The best out-of-sample prediction employ linear regression with LASSO, and both traditional and alternative data, yielding

⁴ Our training set comprises fewer observations than the testing dataset. This intentional design mirrors real-world scenarios, in which an insurance company often relies on limited existing consumer data to predict health risks for a large pool of potential consumers. Furthermore, limiting the training sample's size fit the LASSO approach, characterized by a small number of observations, but a large number of predictors.

⁵ In a robustness check, we employed logistic regression for the *CI_Claim* prediction, yielding similar results.

Table 1
Summary statistics.

	Sample A					Sample B				
	Mean	S.D	Min	P50	Max	Mean	S.D	Min	P50	Max
Dependent Variables										
CI_Claim	0.00283	0.05	0	0	1					
Medical_Claim_Frequency						1.49555	5.90	0	0	134
Independent Variables										
1. Demographic (2)										
Age	33.21676	6.61	18	32	60	33.52481	6.70	18	33	60
Male	0.45101	0.50	0	0	1	0.46559	0.50	0	0	1
2. Insurance plan (5)										
ln_Max_Insurance_Amountt	12.64422	0.46	9	13	14	12.61493	0.48	9	13	14
Add_Applicant_Waiver	0.01478	0.12	0	0	1	0.01579	0.12	0	0	1
Add_Whole_Life	0.00052	0.02	0	0	1	0.00031	0.02	0	0	1
Plans_Number	1.06087	0.28	1	1	8	1.06856	0.29	1	1	8
Total_Application	2.20504	3.68	1	1	258	2.41321	4.28	1	1	299
3. Applicant-related (2)										
Applicant_Age	33.51033	6.55	18	32	72	33.76767	6.60	18	33	67
Applicant_Male	0.36428	0.48	0	0	1	0.37577	0.48	0	0	1
4. Additional demographic (1)										
Children dummies										
Child	0.18308	0.39	0	0	1	0.19801	0.40	0	0	1
Ukw_Child	0.75351	0.43	0	1	1	0.73437	0.44	0	1	1
No_Child	0.06341	0.24	0	0	1	0.06762	0.25	0	0	1
5. Occupation history (1)										
Occupation dummies										
Programmer	0.00328	0.06	0	0	1	0.00353	0.06	0	0	1
Business_Owner	0.00805	0.09	0	0	1	0.00801	0.09	0	0	1
Corporate_Personnel	0.11112	0.31	0	0	1	0.11947	0.32	0	0	1
Accountant	0.00499	0.07	0	0	1	0.00448	0.07	0	0	1
Insurance_Agent	0.01168	0.11	0	0	1	0.01657	0.13	0	0	1
Public_Servant	0.01062	0.10	0	0	1	0.01157	0.11	0	0	1
Other_Occupation	0.04771	0.21	0	0	1	0.04999	0.22	0	0	1
Doctor	0.00149	0.04	0	0	1	0.00195	0.04	0	0	1
Driver	0.00513	0.07	0	0	1	0.00723	0.08	0	0	1
Building_Worker	0.00938	0.10	0	0	1	0.01204	0.11	0	0	1
Lawyer	0.00061	0.02	0	0	1	0.00075	0.03	0	0	1
Coach	0.00002	0.00	0	0	1	0.00003	0.01	0	0	1
Airport_Staff	0.00007	0.01	0	0	1	0.00008	0.01	0	0	1
Bus_Staff	0.00054	0.02	0	0	1	0.00083	0.03	0	0	1
Deliveryman	0.00106	0.03	0	0	1	0.00222	0.05	0	0	1
Farmer	0.00094	0.03	0	0	1	0.00100	0.03	0	0	1
Port_Staff	0.00072	0.03	0	0	1	0.00061	0.02	0	0	1
Train_Staff	0.00004	0.01	0	0	1	0.00003	0.01	0	0	1
White_Collar	0.05650	0.23	0	0	1	0.05766	0.23	0	0	1
Researcher	0.00002	0.00	0	0	1	0.00003	0.01	0	0	1
Teacher	0.00310	0.06	0	0	1	0.00309	0.06	0	0	1
Police_Officer	0.00185	0.04	0	0	1	0.00178	0.04	0	0	1
Retiree	0.00101	0.03	0	0	1	0.00111	0.03	0	0	1
Bank_Staff	0.00990	0.10	0	0	1	0.01051	0.10	0	0	1
Sales	0.00810	0.09	0	0	1	0.00867	0.09	0	0	1
Ukw_Occupation	0.70457	0.46	0	1	1	0.67985	0.47	0	1	1
6. Device used (3)										
Device_Brand dummies										
Huawei	0.18983	0.39	0	0	1	0.20016	0.40	0	0	1
Xiaomi	0.05923	0.24	0	0	1	0.06236	0.24	0	0	1
Oppo	0.05221	0.22	0	0	1	0.05625	0.23	0	0	1
Vivo	0.05482	0.23	0	0	1	0.06089	0.24	0	0	1
Samsung	0.02201	0.15	0	0	1	0.02394	0.15	0	0	1
Other_Brand	0.04905	0.22	0	0	1	0.05135	0.22	0	0	1
Ukw_Brand	0.57680	0.49	0	1	1	0.54953	0.50	0	1	1
Telecom_Company dummies										
China_Telecom	0.05853	0.23	0	0	1	0.06548	0.25	0	0	1
China_Mobile	0.21920	0.41	0	0	1	0.22429	0.42	0	0	1
China_Unicom	0.07288	0.26	0	0	1	0.08449	0.28	0	0	1
Other_Telecom_Company	0.00151	0.04	0	0	1	0.00167	0.04	0	0	1
Ukw_Telecom_Company	0.64969	0.48	0	1	1	0.62635	0.48	0	1	1
Device Type dummies										
Cellphone	0.41980	0.49	0	0	1	0.44666	0.50	0	0	1
Tablet	0.00081	0.03	0	0	1	0.00106	0.03	0	0	1
Ukw_Device_Type	0.57945	0.49	0	1	1	0.55231	0.50	0	1	1
7. Borrowing history (2)										
Receive_Borrow_Notification	0.00103	0.03	0	0	1	0.00150	0.04	0	0	1
Borrow_AppNumber	0.01591	0.19	0	0	9	0.01852	0.21	0	0	8

(continued on next page)

Table 1 (continued)

	Sample A					Sample B				
	Mean	S.D	Min	P50	Max	Mean	S.D	Min	P50	Max
8. Investment (2)										
Money_Management_AppNumber	0.11908	0.51	0	0	9	0.12645	0.52	0	0	8
Receive_Investment_Notification	0.00045	0.02	0	0	1	0.00050	0.02	0	0	1
9. Travel experience (3)										
Receive_Driver_Notification	0.00418	0.06	0	0	1	0.00706	0.08	0	0	1
Number_Cities_Visit	0.07271	0.38	0	0	5	0.07626	0.39	0	0	5
Number_Provinces_Visit	0.07624	0.41	0	0	9	0.08032	0.42	0	0	8
10. Occupation from text and apps (6)										
Teacher_AppNumber	0.00130	0.05	0	0	8	0.00128	0.04	0	0	2
Doctor_AppNumber	0.00221	0.09	0	0	7	0.00209	0.09	0	0	7
Deliver_AppNumber	0.00418	0.09	0	0	5	0.00489	0.09	0	0	4
Receive_InsAgent_Notification	0.00677	0.08	0	0	1	0.00879	0.09	0	0	1
Freight_AppNumber	0.00202	0.06	0	0	6	0.00186	0.05	0	0	3
Medical_AppNumber	0.03810	0.24	0	0	9	0.04134	0.24	0	0	5
11. Living characteristic (3)										
Fit_AppNumber	0.06393	0.35	0	0	8	0.07120	0.37	0	0	7
Receive_Lottery_Notification	0.07156	0.26	0	0	1	0.07215	0.26	0	0	1
Use_Cellphone_inNight	0.21985	0.41	0	0	1	0.22257	0.42	0	0	1
12. Parenthood (2)										
Maternal_AppNumber	0.07721	0.46	0	0	9	0.08466	0.47	0	0	9
Child_Education_AppNumber	0.07050	0.43	0	0	9	0.07690	0.44	0	0	9
13. Insurance purchase (1)										
Receive_Ins_Notification	0.00214	0.05	0	0	1	0.00286	0.05	0	0	1
14. Number of financial institutions inquired (5)										
Financial_Inst_Number30	0.02369	0.21	0	0	9	0.02725	0.24	0	0	10
Financial_Inst_Number60	0.04276	0.33	0	0	12	0.04771	0.37	0	0	14
Financial_Inst_Number90	0.05968	0.42	0	0	13	0.06751	0.47	0	0	16
Financial_Inst_Number180	0.13911	0.72	0	0	19	0.15600	0.81	0	0	21
Financial_Inst_Number240	0.21551	0.99	0	0	40	0.24172	1.14	0	0	40
15. P2p institutions check time (5)										
P2p_institution_Checktime30	0.00308	0.12	0	0	15	0.00350	0.11	0	0	11
P2p_institution_Checktime60	0.00679	0.20	0	0	20	0.00765	0.21	0	0	17
P2p_institution_Checktime90	0.00965	0.25	0	0	24	0.01159	0.28	0	0	24
P2p_institution_Checktime180	0.01955	0.42	0	0	32	0.02427	0.46	0	0	32
P2p_institution_Checktime240	0.02788	0.54	0	0	44	0.03495	0.62	0	0	44
16. Installment loan check time (5)										
Installment_Loan_Checktime30	0.00043	0.03	0	0	3	0.00039	0.03	0	0	3
Installment_Loan_Checktime60	0.00058	0.04	0	0	3	0.00064	0.04	0	0	3
Installment_Loan_Checktime90	0.00090	0.05	0	0	3	0.00111	0.05	0	0	3
Installment_Loan_Checktime180	0.00526	0.12	0	0	9	0.00537	0.12	0	0	7
Installment_Loan_Checktime240	0.01150	0.23	0	0	11	0.01054	0.22	0	0	11
17. Payday loan checktime (5)										
Payday_Loan_Checktime30	0.00196	0.06	0	0	4	0.00206	0.07	0	0	3
Payday_Loan_Checktime60	0.00477	0.13	0	0	8	0.00464	0.13	0	0	8
Payday_Loan_Checktime90	0.00954	0.22	0	0	10	0.00970	0.21	0	0	12
Payday_Loan_Checktime180	0.02264	0.42	0	0	20	0.02600	0.45	0	0	23
Payday_Loan_Checktime240	0.03508	0.61	0	0	34	0.04290	0.68	0	0	38
18. Other financial institutions check time (5)										
Other_Fin_InstChecktime30	0.03674	0.46	0	0	26	0.08636	0.52	0	0	26
Other_Fin_InstChecktime60	0.07630	0.79	0	0	60	0.12965	0.86	0	0	60
Other_Fin_InstChecktime90	0.11501	1.07	0	0	77	0.27567	1.16	0	0	77
Other_Fin_InstChecktime180	0.24168	1.88	0	0	108	0.40776	2.07	0	0	108
Other_Fin_InstChecktime240	0.35616	2.46	0	0	108	0.08636	2.78	0	0	108
19. Financial institutions check time (5)										
Financial_InstChecktime30	0.06740	0.69	0	0	38	0.07955	0.79	0	0	30
Financial_InstChecktime60	0.13454	1.19	0	0	60	0.15272	1.35	0	0	60
Financial_InstChecktime90	0.20221	1.65	0	0	77	0.22999	1.85	0	0	77
Financial_InstChecktime180	0.46425	3.11	0	0	144	0.52784	3.58	0	0	151
Financial_InstChecktime240	0.68909	4.46	0	0	234	0.78619	5.25	0	0	234
Number of Observations	55,560					35,967				

Notes: This table presents the summary statistics of Sample A and Sample B, respectively. We construct Sample A when using the *CI_Claim* as the dependent variable, which is a binary variable that equals 1 if an individual filed a critical illness insurance claim before March 2022 and zero if not. We construct Sample B when using the *Medical_Claim_Frequency* as the dependent variable, which is the total number of claims of the public health insurance (i.e. the basic medical insurance) two years before an individual's critical illness insurance policy application. Sample B includes only individuals whose insured provinces are among the 13 provinces where we have access to the public health insurance claim data. We categorize 63 independent variables into 19 groups. We define age and gender as traditional data, which are in group (1) and the rest 61 variables in groups (2) to (19) as alternative data. The numbers of variables in each group are presented in the parenthesis after the name of the group. Please refer to the section of Sample construction and summary statistics for details

an average out-of-sample R^2 of 0.0492, a 151 % improvement in prediction accuracy. Decomposing this improvement, big data contributes 96 %, while the LASSO model contributes 4 %.

In Column (3) of Table 2, we assess whether alternative data provide additional predictive power conditional on past medical history. We incorporate *Medical_Claim_Frequency* as a proxy for past medical history in the traditional data. We restrict Sample A to

Table 2
Performance of out-of-sample predictions.

Model	Variables	(1) CI_Claim (in AUC)	(2) ln(1 + Medical_ Claim_Frequency) (in out-of-sample R^2)	(3) CI_Claim (in AUC)
Sample		A	B	A∩B
(1) OLS	Traditional_data	0.622	0.0196	0.566
(2) OLS	Traditional_data + Medical_Claim_Frequency	-	-	0.572
(3) OLS	Traditional_data + Alternative_data	0.673	0.048	0.582
(4) OLS	Traditional_data + Medical_Claim_Frequency + Alternative_data	-	-	0.589
(5) LASSO	Traditional_data + Alternative_data	0.696	0.0492	0.613
(6) LASSO	Traditional_data + Medical_Claim_Frequency + Alternative_data	-	-	0.614
(7) Overall improvement		61 %	151 %	58 %
(8) Big Data contribution		69 %	96 %	40 %
(9) Model contribution		31 %	4 %	60 %
Number of observations		55,560	35,967	28,310

Notes: [Table 2](#) presents out-of-sample prediction results where we define *Age*, *Age*², *Age*³ and gender as traditional data, and the other 61 variables as alternative data. Column (1) uses *CI_Claim* as dependent variable with Sample A. Column (2) uses *ln(1 + Medical_Claim_Frequency)* as dependent variable with Sample B. The sample in Column (3) is the intersection of Sample A and Sample B, where we include *Medical_Claim_Frequency* as a independent variable in [Eq. \(3\)](#). We randomly separate the samples into the training sample (30 % observations) and the testing sample (70 % observations). We repeat the above procedure for 100 times and report the mean values in Rows (1) to (6). Row (7) reports the overall improvement of AUC (taking 0.5 as the benchmark, $(0.696-0.5)/(0.622-0.5)-1 = 61\%$) and out-of-sample R^2 ($0.0492/0.0196$)-1 = 151 %. The decomposition of the contributions of Big Data and the Model can be calculated similarly through the results from Row (1) to Row (6).

individuals living in the 13 provinces where we have access to public health insurance claim data, creating a nexus of Samples A and B. In this restricted sample, the baseline out-of-sample prediction using OLS and predictors of traditional data, as well as *Medical_Claim_Frequency*, yields an average AUC of 0.572. The best out-of-sample prediction employ LASSO, traditional data, *Medical_Claim_Frequency*, and alternative data, resulting in an average AUC of 0.614, a 58 % improvement in predictive accuracy. OLS with traditional data, *Medical_Claim_Frequency*, and alternative data achieve an average AUC of 0.589. Alternative data's contribution is 40 % in predictive power, while the LASSO model contribute 60 %.⁶

The relatively low AUC and out-of-sample R^2 throughout the specifications, even after entering alternative data, are not surprising. Due to CI claims' rare-event nature, identifying claimants is by no means an easy task. *Medical_Claim_Frequency* also has many zero values and exhibits extreme right-skewed distribution (e.g., Tweedie distribution), making it difficult to obtain an ideal out-of-sample prediction using linear models. From another perspective, the difficulties in predicting health risks also justify the value of health insurance.

To examine further whether alternative data provide additional information beyond the demographic variables, we present the correlation matrix of all alternative data variables with age and gender in [Appendix Table A3](#). As shown in this table, although certain alternative data variables exhibit a some degree of correlation with age or gender (mostly less than 0.1), a notable portion demonstrate insignificant correlation coefficients. This suggests that a significant number of our big data predictors provide distinct and potentially valuable new signals for insurance risk assessment beyond traditional demographics.

LASSO selection results and predictors' relative importance

[Fig. 1](#) presents the LASSO selection results. We derive the coefficients by using LASSO to estimate [Eq. \(3\)](#), and we include fixed-effects insured provinces, application months, and application years. Among 63 traditional and alternative data variables, in Column (1), in which the dependent variable is *CI_Claim*, 23 variables are selected, and 22 risk predictors were alternative data. The five most important risk predictors are *Plans_Number* (number of insurance products applied), *Age*³, *Occupation*, *Device_Brand*, and *Add_Whole_Life* (whether the insured applied for additional whole life insurance). In Column (2), in which the dependent variable is *Medical_Claim_Frequency*, 38 variables are selected, and 36 risk predictors are alternative data. The five most important risk predictors were *Age* (and *Age*³), *Male*, *Plans_Number* (the number of insurance products applied), *Total_Application* (the number of insurance applications made for the insured), and *ln_Max_Insurance_Amount* (the logarithm of the maximum insurance amount applied for the insured).

[Fig. 1](#) presents the LASSO-selected predictors, with the bars' lengths representing estimated coefficients, demonstrating each selected predictor's relative importance in risk prediction. To derive these, all predictors are standardized to have a mean of 0 and a standard deviation of 1 for prediction power comparison. We also rescale each estimate by dividing them by the maximum absolute value of all estimates, so that the rescaled estimate ranged from -1 to 1. We find that traditional data, namely age and gender, have a large correlation with health risk, following existing industry practice.⁷ LASSO-selected risk predictors also have correlations with health risk conditional on traditional data, revealing unpriced and nontrivial health risk heterogeneity, as well as big data's value.

[Fig. 2](#) presents the Adaptive Group LASSO selection results and estimated coefficients. We used Adaptive Group LASSO to select predictors from 19 groups of variables. In Column (1), in which the dependent variable is *CI_Claim*, five groups of predictors are

⁶ Note that AUC values in Columns (1) and (3) of [Table 2](#) are not comparable due to different samples.

⁷ Although *Age*³ has opposite effects on the two health risk proxies, it does not mean that the effects of age have opposite effects because the impact of age on health risk is nonlinear, estimated as $\hat{\beta} \times (Age, Age^2, Age^3)'$. Online [Appendix Figure A1](#) illustrates age's impact in detail.

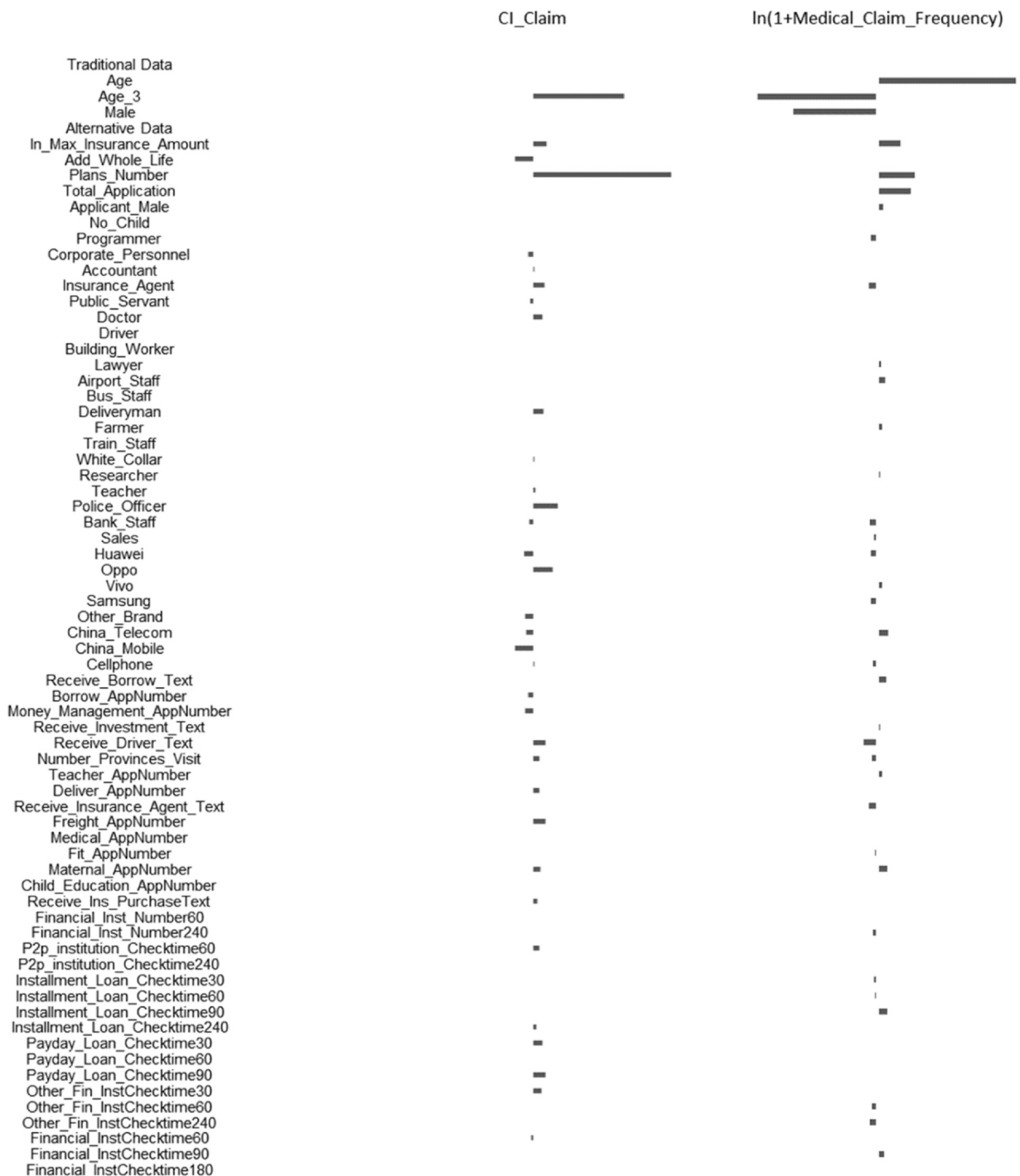


Fig. 1. Rescaled LASSO coefficients. Notes: This figure plots rescaled coefficients, range from -1 to 1 , of LASSO-selected risk predictors. All covariates are standardized to have a mean of 0 and a standard deviation of 1. We also rescale each coefficient by dividing them by the maximum absolute value of respective coefficients. The bar length therefore shows the relative importance of each selected covariate in risk prediction. In Column (1), 23 variables are selected from 63 variables, and 22 of them are alternative data. In Column (2), 38 variables are selected from 63 variables, and 36 variables are alternative data.

selected: demographic; insurance plan; device used; travel experience; and payday loan check time. In Column (2), in which the dependent variable is *Medical_Claim_Frequency*, three groups were selected: demographic; insurance plan; and travel experience. Thus, the most fruitful data collection areas are digital device used, recent travel experience, and insureds' credit records, in addition to traditional demographic and insurance policy information.

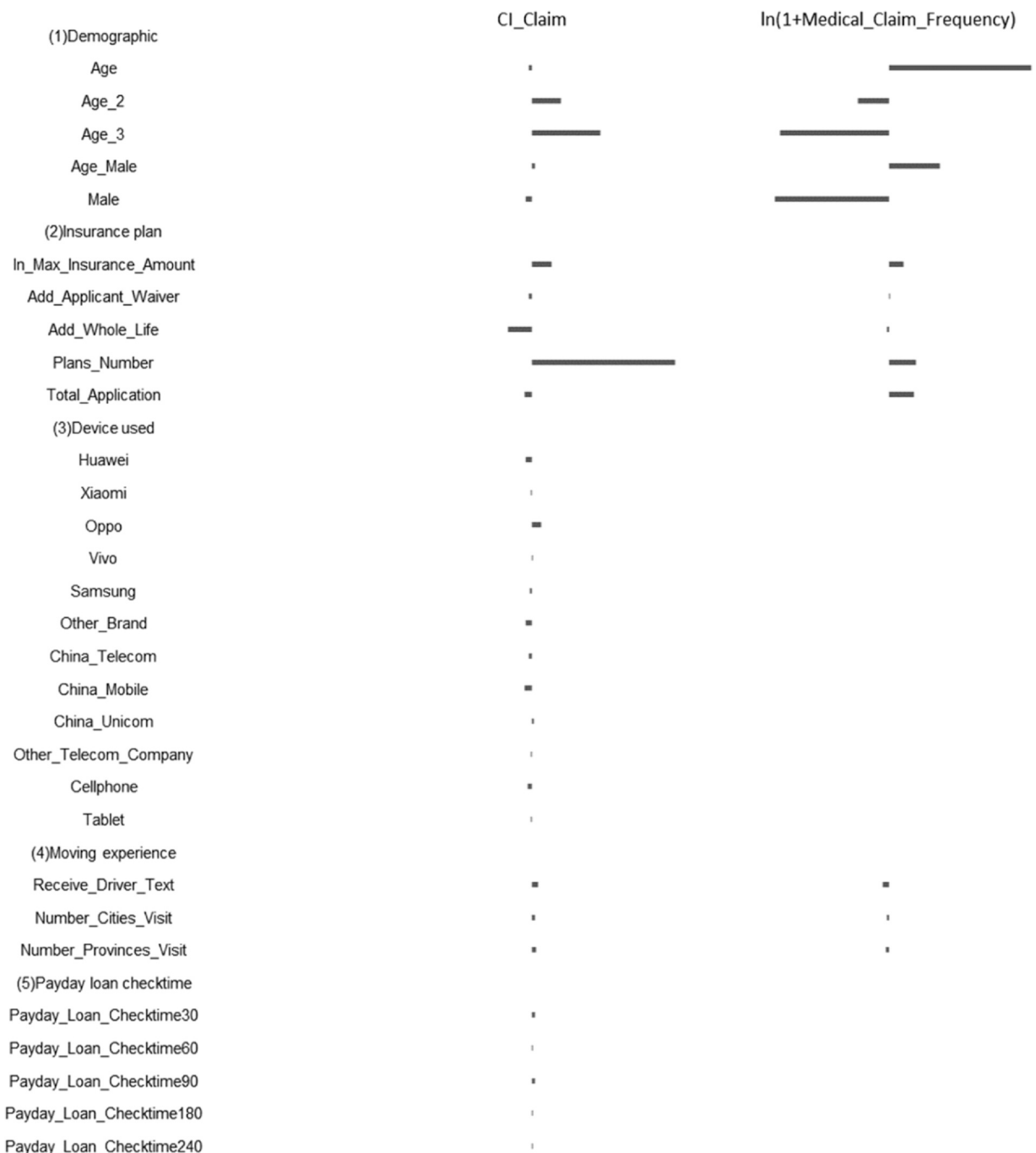


Fig. 2. Rescaled Adaptive Group LASSO coefficients. Note: This figure plots rescaled coefficients of Adaptive Group LASSO-selected predictors individually. All covariates are standardized to have a mean of 0 and a standard deviation of 1 for the prediction power comparison. We also rescale each coefficient by dividing them by the maximum absolute value of respective coefficients, so the rescaled coefficients range from -1 to 1 . The bar length therefore shows the relative importance of each selected covariate in risk prediction. In Column (1), five groups of predictors are selected among 19 groups: Demographic, Insurance plan, Device used, Travel experience, and Payday loan check time. In Column (2), three groups are selected among 19 groups: Demographic, Insurance plan, and Travel experience.

Economic interpretation

LASSO is more transparent compared with other black-box machine learning methods. Table 3 presents Post-LASSO OLS coefficients. Column (1) used *CI_Claim* as the dependent variable, and Column (2) use $\ln(1 + \text{Medical_Claim_Frequency})$. We use LASSO to select risk predictors from traditional and alternative data, the same as in Fig. 1. The Post-LASSO OLS coefficients are obtained by regressing risk proxies on LASSO-selected risk predictors with fixed effects.

Table 3
Post-LASSO OLS on selected risk predictors.

VARIABLES	(1)CI_Claim		(2)ln(1 + Medical_Claim_Frequency)	
	Coef.	S.E.	Coef.	S.E.
Age			0.023***	(0.002)
Age_3	0.0000001***	(0.000)	−0.0000052***	(0.000)
Male			−0.137***	(0.009)
ln_Max_Insurance_Amount	0.001**	(0.001)	0.032***	(0.009)
Add_Whole_Life	−0.023***	(0.007)	−0.125	(0.206)
Plans_Number	0.012***	(0.003)	0.106***	(0.016)
Total_Application			0.006***	(0.001)
Applicant_Male			0.016*	(0.009)
No_Child			−0.006	(0.015)
Programmer			−0.121**	(0.061)
Corporate_Personnel	−0.001	(0.001)	−0.011	(0.012)
Accountant	−0.002***	(0.000)		
Insurance_Agent	0.003	(0.003)	−0.066**	(0.029)
Public_Servant	−0.002***	(0.000)	−0.029	(0.034)
Doctor	0.009	(0.012)		
Driver			−0.030	(0.043)
Building_Worker			−0.023	(0.033)
Lawyer			0.151	(0.131)
Airport_Staff			0.744*	(0.392)
Bus_Staff			−0.081	(0.124)
Deliveryman	0.010	(0.014)		
Farmer			0.152	(0.113)
Train_Staff			−0.293	(0.678)
White_Collar	−0.001	(0.001)		
Researcher			0.634	(0.678)
Teacher	0.003	(0.006)	−0.040	(0.065)
Police_Officer	0.016	(0.013)		
Bank_Staff	−0.002***	(0.000)	−0.079**	(0.036)
Sales			−0.049	(0.039)
Huawei	−0.001	(0.001)	−0.013	(0.013)
Oppo	0.003*	(0.001)	0.015	(0.018)
Vivo			0.026	(0.018)
Samsung			−0.045*	(0.025)
Other_Brand	−0.001**	(0.001)		
China_Telecom	−0.001	(0.001)	0.042***	(0.015)
China_Mobile	−0.001**	(0.001)		
Cellphone	−0.0000243	(0.001)	−0.005	(0.015)
Receive_Borrow_Text			0.195**	(0.094)
Borrow_AppNumber	−0.001***	(0.000)		
Money_Management_AppNumber	−0.001***	(0.000)	−0.004	(0.007)
Receive_Investment_Text			0.140	(0.160)
Receive_Driver_Text	0.007	(0.006)	−0.155***	(0.043)
Number_Provces_Visit	0.001	(0.001)	−0.015*	(0.009)
Teacher_AppNumber			0.112	(0.084)
Deliver_AppNumber	0.003	(0.005)	−0.031	(0.039)
Receive_Insurance_Agent_Text			−0.101**	(0.042)
Freight_AppNumber	0.006	(0.007)	0.056	(0.068)
Medical_AppNumber			−0.008	(0.016)
Fit_AppNumber			−0.008	(0.010)
Maternal_AppNumber	0.001	(0.001)	0.020**	(0.008)
Child_Education_AppNumber			−0.007	(0.009)
Receive_Ins_PurchaseText	0.005	(0.008)	0.055	(0.072)
Financial_Inst_Number60			−0.012	(0.020)
Financial_Inst_Number240			−0.008	(0.007)
P2p_institution_Checktime60	0.002	(0.002)		
P2p_institution_Checktime240			0.003	(0.008)
Installment_Loan_Checktime30			−0.130	(0.183)
Installment_Loan_Checktime60			−0.175	(0.178)
Installment_Loan_Checktime90			0.239**	(0.101)
Installment_Loan_Checktime240	0.001	(0.001)	−0.017	(0.018)
Payday_Loan_Checktime30	0.005	(0.005)		
Payday_Loan_Checktime60			−0.032	(0.032)
Payday_Loan_Checktime90	0.003	(0.002)		
Other_Fin_InstChecktime30	0.002*	(0.001)		
Other_Fin_InstChecktime60			−0.007	(0.008)
Other_Fin_InstChecktime240			−0.005	(0.003)
Financial_InstChecktime60	−0.001**	(0.000)		
Financial_InstChecktime90			0.004	(0.006)
Financial_InstChecktime180			0.004	(0.003)
Financial_InstChecktime240	−0.0001283*	(0.000)		
Constant	−0.033***	(0.009)	−0.430***	(0.119)

(continued on next page)

Table 3 (continued)

VARIABLES	(1)CI_Claim		(2)ln(1 + Medical_Claim_Frequency)	
	Coef.	S.E.	Coef.	S.E.
Observations	55,560		35,967	
Insured province FE	Yes		Yes	
Months of application FE	Yes		Yes	
Years of application FE	Yes		Yes	
Number of Vars. selected	23		38	
Adjusted. R-squared			0.238	
AUC	0.784			

Notes: This table presents the results of OLS estimation using variables selected by the LASSO. Column (1) uses *CI_Claim* as the dependent variable with 23 selected risk predictors as the independent variables. Column (2) uses *ln(1 + Medical_Claim_Frequency)* as the dependent variable with 38 selected risk predictors as the independent variables. The impact of age is nonlinear, which is discussed in Figure A1. Robust standard error are in parentheses. *, **, and *** indicate statistical significance at 10 %, 5 %, and 1 % levels, respectively.

The results indicate that older people are at higher risk. Males ages 18–60 are at lower health risk than females. Individuals with higher insurance demand might be associated with higher risk: A 1 % increase in the maximum insurance amount applied implies a 0.14 % higher probability of filing a CI claim, and one more different insurance product applied implies a 1.2 % higher probability of filing a CI claim. This is a clear indication of adverse selection.

Table 4

Post-adaptive group LASSO OLS on selected variables.

VARIABLES	(1)CI_Claim		(2)ln(1 + ln_Medical_Claim_Frequency)	
	Coef.	S.E.	Coef.	S.E.
Age	0.005**	(0.002)	0.140***	(0.020)
Age_2	−0.0001**	(0.000)	−0.003***	(0.001)
Age_3	0.000002**	(0.000)	0.00003***	(0.000)
Age_Male	0.00003	(0.000)	0.004***	(0.001)
Male	−0.001	(0.003)	−0.260***	(0.037)
ln_Max_Insurance_Amount	0.002***	(0.001)	0.031***	(0.009)
Add_Applicant_Waiver	−0.001	(0.002)	−0.006	(0.029)
Add_Whole_Life	−0.029***	(0.007)	−0.133	(0.206)
Plans_Number	0.012***	(0.003)	0.104***	(0.016)
Total_Application	−0.0001	(0.000)	0.006***	(0.001)
Huawei	0.001	(0.004)		
Xiaomi	0.002	(0.004)		
Oppo	0.005	(0.004)		
Vivo	0.003	(0.004)		
Samsung	0.002	(0.004)		
Other_Brand	0.001	(0.004)		
China_Telecom	−0.001	(0.001)		
China_Mobile	−0.001	(0.001)		
China_Unicom	0.00001	(0.001)		
Other_Telecom_Company	−0.002**	(0.001)		
Cellphone	−0.002	(0.004)		
Tablet	−0.004	(0.004)		
Receive_Driver_Text	0.007	(0.006)	−0.159***	(0.043)
Number_Cities_Visit	−0.003	(0.005)	0.0224	(0.044)
Number_Provinces_Visit	0.003	(0.005)	−0.036	(0.040)
Payday_Loan_Checktime30	0.009	(0.009)		
Payday_Loan_Checktime60	−0.006	(0.006)		
Payday_Loan_Checktime90	0.007	(0.007)		
Payday_Loan_Checktime180	−0.001	(0.001)		
Payday_Loan_Checktime240	−0.001	(0.001)		
Constant	−0.085***	(0.025)	−1.69***	(0.253)
Observations	55,560		35,967	
Insured province FE	Yes		Yes	
Months of application FE	Yes		Yes	
Years of application FE	Yes		Yes	
Number of groups selected	5		3	
Adjusted. R ²			0.238	
AUC	0.782			

Notes: This table presents the results of OLS estimation on the variables selected by the Adaptive Group LASSO. Column (1) uses *CI_Claim* as the dependent variable, and Column (2) uses *ln(1 + Medical_Claim_Frequency)* as the dependent variable. In Column (1), five groups of predictors are selected among 19 groups: Demographic, Insurance plan, Device used, Travel experience, and Payday loan check time. In Column (2), three groups are selected among 19 groups: Demographic, Insurance plan, and Travel experience. Robust standard error are in parentheses. *, **, and *** indicate statistical significance at 10 %, 5 %, and 1 % levels, respectively.

Table 5
Robustness checks for alternative models.

Measurement: AUC		Dependent Variable: <i>CI_Claim</i>	
Method to deal with imbalanced data		(1) None	(2) Weighting
Logistic	Tradition data	0.640	0.633
Logistic	Tradition data + Alternative data	0.612	0.605
LASSO	Tradition data + Alternative data	0.694	0.685
Overall improvement		39 %	39 %
Number of observations		55,560	55,560

Notes: This table present the contribution of Big Data and LASSO by using alternative models. Column (1) uses the standard logistic regression; Column (2) assigns different weights to observations, where the weight is the inverse of proportion of claimants and non-claimants frequencies. The measure for prediction accuracy is AUC.

People with higher socioeconomic status might be healthier (Fiscella & Williams, 2004). People with better working conditions (i.e., corporate personnel, public servants, white collar workers, bank staff) are healthier, and Huawei and Samsung (high-end brands) users are healthier than other Android users, while Vivo and Oppo (low-end brands) users are at higher health risk. Travel experience might reflect an insured's lifestyle, e.g., one might be too busy to go to hospitals, putting them at higher health risk. Notably, we do not offer causal interpretations, but rather suggestive economic understandings of the predictions.

Table 4 presents Post-Adaptive Group LASSO coefficients, which were derived by regressing health risk proxies on Adaptive Group LASSO-selected groups of risk predictors. Notably, the regressors are different from those selected by LASSO because variables are selected in a group manner. In addition to those presented in Table 3, the number of credit history checks conducted by payday loan institutions, an indication of unhealthy financial condition, notably is related positively to health risk (Bialowolski et al., 2021).

Robustness checks

Alternative models

Table 5 presents out-of-sample predictions by logistic regression, with *CI_Claim* as the dependent variable. Column (1) presents the results from standard logistic regression, while Column (2) presents results that assign different weights to observations, in which the weight is the inverse of the proportion of claimants and non-claimants' frequencies. Both specifications illustrate that alternative data with LASSO improve prediction accuracy. Interestingly, when we include all predictors without LASSO in the model, we encounter an overfitting problem, in which out-of-sample prediction accuracy decrease in both columns compared with the benchmark model, justifying our choice of a linear model with LASSO in our main analyses.

Alternative risk proxies and subsamples

To alleviate concerns over different observation windows of *CI_Claim*, we excluded insureds who filed a claim after one year of the policy's inception. Accordingly, out-of-sample prediction results are presented in Panel A of Table 6, in which all insureds have the

Table 6
Robustness checks for alternative risk proxy and samples.

Panel A: Alternative risk proxy			
Model	Variables	CI_Claim_1year(1-year observation window, in AUC)	
OLS	Traditional data	0.611	
OLS	Traditional data + alternative data	0.663	
LASSO	Traditional data + alternative data	0.688	
Overall improvement		69 %	
Number of observations		55,560	
Panel B: Alternative sample restriction			
Model	Variables	(1)	(2)
		CI_Claim(in AUC)	ln(1 + Medical_Claim_Frequency)(in out-of-sample R ²)
OLS	Traditional data	0.624	0.02
OLS	Traditional data + alternative data	0.669	0.052
LASSO	Traditional data + alternative data	0.698	0.054
Overall improvement		60 %	170 %
Number of observations		53,034	32,586

Notes: Panel A documents the robustness check for the results in Column (1) of Table 2 using an alternative health risk proxy *CI_Claim_1year*, where we exclude insureds who filed a claim one year after the critical illness policy inception. Thus, all individuals have the same observation period for critical illness claims. Panel B documents the results excluding individuals whose insured province is Hubei, the epicenter in China, to exclude the potential influence of COVID-19.

same one-year observation window. Another concern is COVID-19 and the associated lockdown, which might have influenced hospitals' behavior and public health insurance claim data. Some individuals might not have had access to hospitals during the lockdown. We address COVID-19's influence by excluding insureds whose insured province is Hubei, the epicenter of COVID-19 in China. The results are presented in Panel B of Table 6. All of our conclusions hold.

Conclusion

In this paper, we use an enriched dataset provided by Insurtech, a Chinese company, and develop appropriate models to predict insurance risks. We show that big data brings in extra predictive power in addition to traditional demographic and insurance policy data. To provide an explicit predictor interpretation and avoid "overfitting," we apply LASSO and Adaptive Group LASSO to filter the most effective predictors and fruitful data collection fields, respectively.

However, questions remain to be addressed in future research. Although adverse selection might be mitigated with new data and prediction methodologies, the welfare loss from reclassification risk might be exacerbated (Handel et al., 2015). This could become a significant issue as insurance companies gain the ability to assess risks more accurately using big data and machine learning methods, thereby justifying future research. Although our big data and model improve insurance risk predictions, a considerable variation in insurance risk remains unexplained. Nonetheless, the model presented in this paper can be used with other alternative datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author Ruo Jia is an Editor for Risk Sciences and was not involved in the editorial review or the decision to publish this article.

Acknowledgements

We thank National Natural Science Foundation of China (No. 72173005), National Social Science Foundation of China (23& ZD178), and Research Seed Fund of the School of Economics, Peking University, for financial support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.risk.2025.100028](https://doi.org/10.1016/j.risk.2025.100028).

References

- Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.
- Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.
- Bialowolski, P., Weziak-Bialowolska, D., Lee, M., Chen, Y., VanderWeele, T., & McNeely, E. (2021). The role of financial conditions for physical and mental health: Evidence from a longitudinal survey and insurance claim data. *Social Science Medicine*, 281, Article 114041.
- Black, K., Skipper, H. D., & Huebner, S. (1994). *Life insurance*. Prentice Hall.
- Boudt, K., Delmarcelle, O., & Ringoot, P. (2025). A news-monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions. *Risk Sciences* Article 100018.
- Browne, M. J., & Kamiya, S. (2012). A theory of the demand for underwriting. *Journal of Risk and Insurance*, 79(2), 335–349.
- Cather, D. A. (2018). Cream skimming: Innovations in insurance risk classification and adverse selection. *Risk Management and Insurance Review*, 21(2), 335–366.
- Finkelstein, A., & Poterba, J. (2004). Adverse selection in insurance markets: Policyholder evidence from the UK annuity market. *Journal of Political Economy*, 112(1), 183–208.
- Fiscella, K., & Williams, D. (2004). Health disparities based on socioeconomic inequities: Implications for urban health care. *Academic Medicine*, 79(12), 1139–1147.
- Fong, J. H. (2015). Beyond age and sex: Enhancing annuity pricing. *Geneva Risk and Insurance Review*, 40(2), 133–170.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2021). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.
- Geyer, A., Kremslehner, D., & Muermann, A. (2020). Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance*, 87(4), 969–995.
- Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, 88(3), 569–589.
- Handel, B., Hendel, I., & Whinston, M. D. (2015). Equilibria in health exchanges: Adverse selection versus reclassification risk. *Econometrica*, 83(4), 1261–1313.
- Powers, M. R., & Xu, J. (2025). A versatility measure for parametric risk models. *Risk Sciences* Article 100017.
- Qu, B., Tian, Y., & Wei, L. (2025). Reducing adverse selection by taking advantage of ML: Evidence from China's Internet auto insurance. *Risk Sciences* Article 100021.
- Rothschild, M., & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, 90(4), 629–649.
- Shi, J., Yao, Y., & Liu, G. (2018). Modeling individual health care expenditures in China: Evidence to assist payment reform in public insurance. *Health Economics*, 27(12), 1945–1962.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Wang, H., & Leng, C. (2008). A note on adaptive group LASSO. *Computational Statistics Data Analysis*, 52(12), 5277–5286.
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.