Issue Brief

Sourcing Protected Class Information in P&C Insurance

JUNE 2022

Key Points

- Property & casualty insurance companies do not typically collect any protected class attributes other than gender. This issue brief discusses potential options to obtaining protected class information on insureds in the P&C insurance industry to test for practices that may disproportionally harm members of any protected classes.

- The issue brief evaluates three general data collection methods—collecting additional data, capturing existing external data, or imputing data—for desirable attributes such as accuracy, efficiency, and the size of the data set.

- While trade-offs exist between different methods of obtaining protected class data, a combination of the methods would allow for the use of existing data and imputation of missing data while exploring new sources of data.

Background

To study the treatment of protected classes covered by the insurance industry, data is required to perform analyses and arrive at an assessment. However, property & casualty (P&C) insurance companies do not typically collect any protected class attributes other than gender, which makes it difficult to assess with certainty how their models are performing based on protected class attributes.

This issue brief discusses potential options to obtaining protected class information on insureds in the P&C insurance industry to test for practices that may disproportionally harm members of any protected classes, which we will refer to as unfair or unlawful discrimination in the remainder of this issue brief. These potential P&C options may not apply to other practice areas (e.g., life or health) because of the differences in products and how protected class information is currently utilized.

Why would insurance stakeholders need protected class data?

Several active legislative and regulatory initiatives would require insurance stakeholders to test for unfair or unlawful discrimination. These proposals include assessing the correlation between protected classes and insurance rating variables, measuring the difference in rates between classes of insureds, and prohibiting the use of certain rating attributes.

A full discussion of the different methods to assess the treatment of protected classes in P&C insurance is beyond the scope of this issue brief. Each method requires data for the purposes of identifying problems and determining solutions. Obtaining protected class data is one of the first steps in addressing unfair or unlawful discrimination in insurance industry practices.

What protected class data would insurance stakeholders need?

There are several protected class characteristics that have been a part of the discussion regarding unfair or unlawful discrimination in P&C insurance. These characteristics include (but are not limited to):

1. Race
2. Ethnicity
3. Nationality
4. Religion
5. Income
6. Sexual orientation
7. Gender identity

Race has been the focus of much recent discussion; however, a process for testing for unfair or unlawful discrimination could potentially be applied to any protected class characteristic. Stakeholders need to develop a standard set of protected class information to collect for analysis to be consistent.

How would insurance stakeholders obtain protected class data?

Given that P&C insurance companies do not typically collect any protected class attributes currently other than gender, new approaches for sourcing this data would need to be developed. These data sources could be developed by collecting additional data, capturing existing external data, or imputing the data.

Below is a discussion of desirable attributes of a protected class data source and an assessment of different methods to obtain data.

Desired attributes of protected class data source

- **Accurate**—Data represents the correct value.
- **Process-Efficient**—Data is readily available to inform active public policy discussions through a process that is efficient to implement.
- **Cost Efficient**—The process to collect and store data is cost-efficient so as not to significantly impact insurance prices.
- **Sizeable**—A large, credible amount of data is readily accessible to insurance stakeholders.
- **Extensive**—Data for a wide array of protected class data fields is available.

Use of Protected Class Data

If insurance stakeholders obtain protected class data, they will need to ensure that protected class data is only used for appropriate purposes. Protected class data could be misused. Therefore, it is important that security controls are in place. Data access would need to be limited only to those individuals requiring the data for specific discrimination testing. Data use would need to be tracked, managed, and disclosed.

Data-Collection Methods

Protected class data can be obtained through three general methods or a combination of these methods:

- **Collect additional data**—Insurance companies can begin to ask insureds for their protected characteristics on a voluntary basis and record this data.
- **Capture existing external data**—An entity can access existing external data sources with known protected class information to be shared with insurance stakeholders (i.e., analogous to a third-party credit bureau, but for protected class characteristics).
- **Impute data**—Protected class data can be imputed on insurance data using predictive models based on information such as ZIP code and surname to estimate certain elements of protected class information.
- **A combination of the above methods**. This approach would allow for the use of existing external data and the imputation of missing data, while exploring collecting additional data.

Evaluation of Data-Collection Methods

Collect additional data

Collecting additional data by asking insureds for their protected class information (whether at time of purchase, time of renewal, or off-cycle survey) could be accurate because information would be coming directly from customers. However, consumers may decline to participate and/or provide inaccurate information because they may suspect this information could affect their insurance coverage or invade their privacy. Further, each company would incur significant expense to implement capturing the data in all existing policy processing systems. Collecting data from insureds would allow for the collection of a wide array of protected class data fields.

Collecting additional data may not be efficient to implement because insurance companies would need to make technology and business process changes to capture the data. This operational disruption would likely result in additional cost being passed on to consumers. In some states, laws would need to be changed to allow insurance companies to collect this data.

Capture existing external data

Capturing existing external data, such as Census data,[1] would enable a sizable amount of data to be available immediately for many protected classes, but not necessarily every protected class (e.g., sexual orientation). Importing existing external data into internal databases would be comparatively lower-cost and require less effort than collecting additional data from insureds. Existing external data could also be the most accurate if it is collected in an unbiased setting (e.g., the U.S. Census).

Capturing existing external data may be difficult to implement. A new data service or entity may need to be created to gather data in a centralized location and make it accessible to others. This need for centralization would delay the availability of the data—an initiative of this nature would likely need to be approved by several parties and require funding. Finally, existing external data such as Census data may be dated, which is a potential concern if individuals self-identify differently over time or if categories for identification change over time or vary between data sources.

---

1 Privacy and confidentiality of U.S. Census data would need to be addressed but is beyond the scope of this issue brief. Privacy could be protected by anonymizing after third-party matching to carrier data.

Impute data

Imputing data would enable a credible amount of data to be available immediately for some protected class characteristics. This method would be the most efficient to implement and the least expensive because insurance companies could impute data onto their policyholder data with a publicly available algorithm. Imputation of racial data is already utilized by agencies such as the Consumer Financial Protection Bureau (CFPB) to ensure that lenders are complying with fair lending laws.[2]

Imputing data would not be as accurate as other data collection methods. According to Rand, "Concordance between self-reported race/ethnicity and BISG [Bayesian indirect surname geocoding] estimates are typically 90-96% accurate for the four largest racial/ethnic groups (Asian/Pacific Islander, Black, Hispanic and White)," but are "less accurate and not recommended for American Indians/Alaskan Natives and Multi-racial persons."[3] Additional testing of imputation methods would be required to ensure imputation methods are appropriate for analyzing property and casualty insurance and are not biased.

Imputing other types of protected class data such as sexual orientation would not be possible using existing imputation methods. Therefore, imputing data would only be possible for a limited number of protected class data fields but could be used as a first step in collecting certain types of protected class information until more permanent solutions can be implemented. Over time, the integration of imputation with the collection of data from insureds could allow for the development of an accurate, comprehensive set of protected class data.

Data-Collection Evaluation—Summary

The data-collection evaluation described above is summarized in the below table as well as in greater detail in the appendix.

Methodology	Accurate	Process-Efficient	Cost-Efficient	Sizeable	Extensive
Collect additional data	◑	○	○	◑	●
Capture existing external data	◑	◑	◑	●	◑
Impute data	◑	●	●	●	○

● = Methodology meets criteria
◑ = Methodology partially meets criteria
○ = Methodology does not meet criteria

2 Consumer Financial Protection Bureau; *Using publicly available information to proxy for unidentified race and ethnicity—A methodology and assessment*; Summer 2014.
3 Rand Corporate; "When Race/Ethnicity Data Are Lacking"; *RAND Health Quarterly*; 2016.

A combination of the above methods would allow for the use of existing data and imputation of missing data, while exploring new sources of data. This would potentially improve the accuracy, size, and extensiveness of the data but could be less process-efficient as insurers would have to manage multiple data methods.

Conclusion

To study the treatment of protected classes in the insurance industry, data is necessary to perform an analysis and arrive at an assessment. Trade-offs exist between different methods of obtaining protected class data. Working toward adopting one or more of these methods could meet short- and long-term goals of safeguarding against unfair or unlawful discrimination occurring in insurance.

Appendix: Table for Evaluation of Data-Collection Methods

Desired Attribute	Collect Additional Data	Capture Existing External Data	Impute Data
Accurate	Creating additional data source by asking insureds for their protected information at time of purchase or time of renewal could be accurate because information would be coming directly from customers. However, consumers may decline to participate and/or provide inaccurate information because they may suspect this information will affect their insurance coverage or invade their privacy.	Existing external data may be the most accurate if it is collected in an unbiased setting like the U.S. Census. However, existing external data may be dated, which is a potential concern if individuals self-identify differently over time.	Imputing data may not be as accurate as other data collection methods. Best-in-class imputation methods of racial/ethnic groups have been found to be about 90-96% accurate but are not recommended for some groups. Additional testing of imputation methods would be required to ensure imputation methods are appropriate for analyzing property and casualty insurance and are not biased.
Process- Efficient	There would be a multiyear lag between the implementation of collecting additional data and the subsequent analysis unless an off-cycle survey was used.	A new data service or entity may need to be created to gather data in a centralized location and make it accessible to others. This would delay the availability of the data—an initiative of this nature would likely need to be approved by several parties and require funding.	This method would be the most efficient to implement because insurance companies could impute data onto their policyholder data with a publicly available algorithm.
Cost- Efficient	Each company would incur significant expense to implement capturing the data through all existing policy processing systems.	Importing existing external data into internal databases requires comparatively low cost and effort on the part of insurance stakeholders.	This method would be the least expensive because insurance companies could impute data onto their policyholder data with a publicly available algorithm.
Sizable	A credible amount of data would not be available immediately because data would be collected on only a portion of new business and renewal business.	Capturing existing external data, such as Census data, would enable a sizable amount of data to be available immediately for some protected class characteristics (notably race).	Imputing data would enable a credible amount of data to be available immediately for some protected class characteristics (notably race).
Extensive	Creating additional data would allow for the collection of any protected class data field.	Capturing existing external data, such as Census data, would enable many protected classes to be analyzed immediately, but not necessarily every protected class.	Imputing protected class data (other than race) would not be possible using existing imputation methods. Therefore, imputing data would only be possible for a limited number of protected class data fields.
Use of Protected Class Data	Regardless of the methods used by insurance stakeholders to obtain protected class data, they will need to ensure that protected class data is only used for appropriate purposes. Protected class data could be misused. Therefore, it is important that security controls are in place. Data access would need to be limited only to those individuals requiring the data for specific discrimination testing. Data use would need to be tracked, managed, and disclosed.		