July 27, 2020

Kris DeFrain, FCAS, MAAA, CPCU
Director of Research and Actuarial Services
National Association of Insurance Commissioners

Sent via email

Re: CASTF Draft – *Regulatory Review of Predictive Models White Paper*

Dear Kris:

I appreciate this opportunity to comment on the Casualty Actuarial and Statistical Task Force (CASTF)'s June 12, 2020, exposure draft containing potential best practices for the Regulatory Review of Predictive Models (RRPM). I note that it appears that the CASTF is working to bring the review process of its white paper to a close. Clearly this is a topic of interest for the Academy's membership. As in our prior two letters,[1,2] we would like to offer just some brief comments.

I would like to revisit a point from my prior letter. Specifically, in Section VII there is a discussion of regulatory best practices. Item 1.b. discusses the need to determine that individual input characteristics and resulting rating factors are related to the expected loss or expense differences in risk. Later in the document, Appendix B, *Information Elements* A.4.b and B.3.d seeks to obtain information as to the rational relationship or rational explanation that predictive data or predictor variables have to the predicted variable. Predictive data or predictor variables that are related to risk of loss (as demonstrated by analysis of historical insurance loss or expense data across the predictors) are key rational relationships. As we consider this, actuaries are guided by Actuarial Standard of Practice (ASOP) No. 12, *Risk Classification* . Within that ASOP, there are several key considerations to guide both regulators and modelers.

The RRPM White Paper provides considerable latitude in its scope. Perhaps this is in keeping with the range of possibilities that new data sources coupled with broad computing power brings to the predictive modeling field. Insurance underwriting has, for decades, been moved in the direction of greater granularity in its use of data and underwriter judgment. All the while, these efforts have facilitated better pricing accuracy and broader availability of insurance products. In

[1] https://www.actuary.org/sites/default/files/files/publications/CASTF_Predictive_Modeling_Comments_20190122.pdf
[2] https://www.actuary.org/sites/default/files/2019-11/CASTF_Academy_Comments_on_RRPM_3rd_exposure.pdf

1850 M Street NW Suite 300 Washington, DC 20036 Telephone 202 223 8196 Facsimile 202 872 1948 www.actuary.org

short, model innovation has many potential benefits to the insurance market. At the same time, there is the potential for modeling to be stretched too far through such innovations. One would hope that application of the RRPM White Paper best practices, used effectively, finds a balance between innovation and control.

New data sources are ever changing, and especially when considered in the context of technological improvements possible for the insurance industry, pose interesting challenges and opportunities. Properly used, these new tools and access to data should lead to expense reductions that ultimately yield lower costs in the insurance system. That said, new data sources require considerable due diligence as they are assimilated into the modeling process. ASOP No. 23, *Data Quality,* is available to guide actuaries as they consider new information sources. We would again hope that a RRPM process will work collaboratively with ASOP No. 23 around new data sources.

Finally, as CASTF moves toward finalizing its recommendations for the RRPM process and thus toward implementation across the various states, I think that it is important to understand the workload challenges that will perhaps result from the new RRPM requirements. Specifically, will the state regulators have the necessary staffing and/or resources to move toward effective implementation?

Thanks once again for allowing this input. The Academy remains available to assist as CASTF moves forward with this.

Sincerely,

Richard Gibson, MAAA, FCAS
Senior Casualty Fellow
American Academy of Actuaries

1850 M Street NW Suite 300 Washington, DC 20036 Telephone 202 223 8196 Facsimile 202 872 1948 www.actuary.org

July 27, 2020

Kris DeFrain, FCAS, MAAA, CPCU
Director, Research and Actuarial Services
National Association of Insurance Commissioners (NAIC)
NAIC Central Office
1100 Walnut Street, Suite 1500 Kansas City, MO 64106-2197

Sent via e-mail at kdefrain@naic.org

RE: CASTF – Regulatory Review of Predictive Models: version 06-12-2020

The American Property Casualty Insurance Association (APCIA)[1] appreciates the opportunity to provide comment on the NAIC Casualty Actuarial and Statistical Task Force (CASTF) exposure draft, dated June 12, 2020, regarding the *Regulatory Review of Predictive Models*.

APCIA is committed to working collaboratively with the NAIC in support of innovation and the effort to leverage the advancements in technology and data analytics to effectively respond to the changing risks and needs of insurance consumers. APCIA appreciates some changes have been made by the drafting group, such as mapping the best practices with the information elements. In reviewing the document as a whole, however, APCIA's conclusion remains that the information collected in Appendix B will lead to significant increase in the length of time and costs for filing approval with limited regulatory benefit. Importantly, several of these information elements are redundant, highly prescriptive, and overly detailed. Approval time is going to be significantly extended and our ability to respond to the needs of the insurance consumer negatively delayed.

Finally, the Key Regulatory Principles identify principles that the best practices are based on to promote a comprehensive and coordinated review of predictive models across states. Principle two indicates that "State regulators will be able to share information to aid companies in getting insurance products to market more quickly across the states." APCIA recognizes that this is not necessarily new for this version of the White Paper but inquires if CASTF could explain what is meant by that particular principle.

We have included a red-line version of Appendix B for your consideration. A description of the APCIA reasoning for each recommended change is provided in the final column of the chart.

Thank you for your consideration of these comments and APCIA is happy to answer any questions that you may have.

Respectfully submitted,

Angela Gleason

APENDIX B – INFORMATION ELEMENTS AND GUIDANCE FOR A REGULATOR TO MEET BEST PRACTICES' OBJECTIVES (WHEN REVIEWING GLMS)

This appendix identifies the information a regulator may need to review a predictive model used by an insurer to support a personal automobile or home insurance rating plan. The list is lengthy but not exhaustive. It is not intended to limit the authority of a regulator to request additional information in support of the model or filed rating plan. Nor is every item on the list intended to be a requirement for every filing. However, the items listed should help guide a regulator to sufficient information that helps determine if the rating plan meets state specific filing and legal requirements.

Documentation of the design and operational details of the model will ensure business continuity and transparency of models used. Documentation should be sufficiently detailed and complete to enable a qualified third party to form a sound judgment on the suitability of the model for the intended purpose. The theory, assumptions, methodologies, software and empirical bases should be explained, as well as the data used in developing and implementing the model. Relevant testing and ongoing performance testing need to be documented. Key model limitations and overrides need to be pointed out so that stakeholders understand the circumstances under which the model does not work effectively. End-user documentation should be provided and key reports using the model results described. Major changes to the model need to be shared with regulators in a timely manner and documented, and IT controls should be in place, such as a record of versions, change control and access to model.[1] APCIA Comment for Consideration: This paragraph describes documentation that needs to be kept on the model, but also adds to the IT controls for model revisions. Since the majority of the models within the review are the pricing models, which need to be filed with the department of insurance, the IT controls are on the rating plans and not necessarily the model version. Additionally, companies are already subject to robust IT controls. It is also unclear on what is meant by the following sentence: "Major changes to the model need to be shared with regulators in a timely manner. . ." If a company refreshes the model and notices an indication changed for a particular attribute, won't the appropriate mechanism for sharing that information be part of a filing to adjust the rating plan? This sentence could be read to suggest a different type of inform being necessary.

Many information elements listed below are probably confidential, proprietary or trade secret and should be treated as such according to state law. Regulators should be aware of their state laws on confidentiality when requesting data from insurers that may be proprietary or trade secret. For example, some proprietary models may have contractual terms (with the insurer) that prevent disclosure to the public. Without clear necessity, exposing this data to additional dissemination may compromise the model's protection.[2]

Though the list of information is long, the insurer should already have internal documentation on the model for more than half of the information listed. The remaining items on the list require either minimal analysis (approximately 25%) or deeper analysis to generate for a regulator (approximately 25%).

The "Level of Importance to the Regulator's Review" is a ranking of information a regulator may need to review is based on the following level criteria:

Level 1 - This information is necessary to begin the review of a predictive model. These data elements pertain to basic information about the type and structure of the model, the data and variables used, the assumptions made, and the goodness of fit. Ideally, this information would be included in the filing documentation with the initial submission of a filing made based on a predictive model.

Level 2 - This information is necessary to continue the review of all but the most basic models; such as those based only on the filer`s internal data and only including variables that are in the filed rating plan. These data elements provide more detailed information about the model and address questions arising from review of the information in Level 1. Insurers concerned with speed to market may also want to include this information in the filing documentation.

Level 3 - This information is necessary to continue the review of a model where concerns have been raised and not resolved based on review of the information in Levels 1 and 2. These data elements address even more detailed aspects of the model. This information does not necessarily need to be included with the initial submission, unless specifically requested in a particular state, as it is typically requested only if the reviewer has concerns that the model may not comply with state laws.

Level 4 - This information is necessary to continue the review of a model where concerns have been raised and not resolved based on the information in Levels 1, 2, and 3. This most granular level of detail is addressing the basic building blocks of the model and does not necessarily need to be included by the filer with the initial submission, unless specifically requested in a particular state. It is typically requested only if the reviewer has serious concerns that the model may produce rates or rating factors that are excessive, inadequate, or unfairly discriminatory.

Lastly, though the best practices presented in this paper will readily be transferrable to review of other predictive models, the information elements presented here might be useful only with deeper adaptations when starting to review different types of predictive models. If the model is not a GLM, some listed items might not apply, for example, not all predictive models generate p-values or F tests. Depending

[1] Model Risk Management: An Overview, the Modeling Section of the Society of Actuaries, Michele, Bourdeau, The Modeling Platform Issue 4, December 2016.
[2] There are some models that are made public by the vendor and would not result in a hindrance of the model's protection.

on the model type, other considerations might be important but are not listed here. When information elements presented in this appendix is applied to lines of business other than personal automobile and home insurance or other type of models, unique considerations may arise, in particular data volume and credibility may be lower for other lines of business. Regulators should be aware of the context in which a predictive model is deployed, the uses to which the model is proposed to be put, and the potential consequences the model may have on the insurer, its customers, and its competitors. This paper does not delve into these possible considerations, but regulators should be prepared to address them as they arise.

A. SELECTING MODEL INPUT

Section	Information Element	Level of Importance to the Regulator's	Comments	APCIA Comments
1. Available Data Sources				
A.1.a	Review the details of sources for both insurance and non-insurance data used as input to the model (only need sources for filed input characteristics included in the filed model).	1	Request details of ~~all~~ data sources, with a proportionate impact on rates, whether internal to the company or from external sources. For insurance experience (policy or claim), determine whether data are aggregated by calendar, accident, fiscal or policy year and when it was last evaluated. For each data source, get a list of ~~all~~ data elements used as input to the model that came from that source recognizing the need for exceptions for the proprietary components when sufficient confidentiality and trade secret protections are not available. ~~For insurance data, get a list all companies whose data is included in the datasets.~~ Request details of any non-insurance data used (customer-provided or other), whether the data was collected by use of a questionnaire/checklist, whether data was voluntarily reported by the applicant, and whether any of the data is subject to the Fair Credit Reporting Act. If the data is from an outside source, find out what steps the insurer has ~~were~~ taken to verify the outside source has processes and procedures in place to assess the data's ~~was~~ accuracy, completeness and ~~un~~biased characteristics in terms of relevant and representative time frame, representative of potential exposures and lacking in obvious correlation to protected classes. Note that reviewing source details should not make a difference when the model is new or refreshed~~; refreshed models would report the prior version list with the incremental changes due to the refresh.~~	Not all Requesting a contributing company list is an unnecessary overreach and, in some cases, will be unknown due to aggregation and anonymization. In addition, there could be contractual restrictions for sharing this information. APCIA respectfully believes that the remaining data elements should be sufficient. At the very least this should not be considered essential level 1 data. Insurance companies can take some verification steps, but most of the work will have to be completed by the third-party. The suggestion for refreshed models infers that a company needs to document changes in data source on a refresh and that is not a requirement. There may need to be some clarification around what is considered "insurance" and non-insurance data. Also, consideration should be given to the application of CAS ASOP 12.

A.1.b	Reconcile aggregated insurance data underlying the model with available external insurance reports.	4	Accuracy of insurance data should be reviewed. It is assumed that the data in the insurer's data banks is subject to routine internal company audits and reconciliation. "Aggregated data" is straight from the insurer's data banks without further modification (e.g., not scrubbed or transformed for the purposes of modeling). In other words, the data would not have been specifically modified for the purpose of model building. The company should provide some form of reasonability check that the data makes sense when checked against other audited sources.	
A.1.c	Review the geographic scope and geographic exposure distribution of the raw data for relevance to the state where the model is filed.	2	~~The company should explain how the data used to build the model makes sense for a specific state~~. The regulator should inquire which states were included in the data underlying the model build, testing and validation. The company should provide an explanation where the data came from geographically and that it is a good representation for a state, i.e., the distribution by state should not introduce a geographic bias. For example, there could be a bias by peril or wind-resistant building codes. Evaluate whether the data is relevant to the loss potential for which it is being used. For example, verify that hurricane data is only used where hurricanes can occur.	Many models are developed using a countrywide or a regional dataset, rarely is a model built for a specific state. This review must balance the question of relevance with potential scarcity of data for a particular state.
2. Sub-Models -				
A.2.a	Consider the relevance of (e.g., is there a bias) of overlapping data or variables used in the model and sub-models.	1	Check if the same variables/datasets were used in both the model, a sub-model or as stand-alone rating characteristics. If so, verify the insurance company has processes and procedures in place to assess and address ~~there was no~~ double-counting or redundancy.	
A.2.b	Determine if the sub-model was previously approved (or accepted) by the regulatory agency.	1	If the sub-model was previously approved, that may reduce the extent of the sub-model's review. If approved, obtain the SERFF number and verify when and that it was the same model currently under review. However, previous approvals do not necessarily confer a guarantee of ongoing approval, for example when statutes and regulations have changed or if a model's indications have been undermined by subsequent empirical experience. However, knowing whether a model has been previously approved can help focus the regulator's efforts and determine whether or not the prior decision needs to be revisited.	Confidentiality and competitiveness issues may arise when disclosing vendor information. Since this is an element for expediting review, it is reasonable to make sure the insurance company consents to the vendor identification and dialogue. In addition, the insurer may not have the right to disclose if a sub-model is purchased. Extending the regulator's review to the underlying sub-model in the same breadth and depth as the insurer's filed model seems very impractical.

A.2.c	Determine if sub-model output was used as input to the GLM; obtain the vendor name, and the name and version of the sub-model.	1	To accelerate the review of the filing, and consented to by the company, get the name and contact information for a representative from the vendor. The company should provide the name of the third-party vendor and a contact in the event the regulator has questions. The "contact" can be an intermediary at the insurer, e.g., a filing specialist, who can place the regulator in direct contact with a Subject Matter Expert (SME) at the vendor." Examples of such sub-models include credit/financial scoring algorithms and household composite score models. ~~Sub-models can be evaluated separately and in the same manner as the primary model under evaluation.~~ A sub-model contact for additional information should be provided. SMEs on sub-model may need to be brought into the conversation with regulators (whether in-house or 3rd-party sub-models are used).	In addition to the recommended changes in the comments, they should clarify that this element is directed at vendor contact information and not the accuracy of the vendor data. If it is vendor data accuracy, then the burden of that question, as noted elsewhere in these comments, will be on the vendor, not the insurer.
A.2.d	If using catastrophe model output, identify the vendor and the model settings/assumptions used when the model was run.	1	For example, it is important to know hurricane model settings for storm surge, demand surge, long/short-term views. To accelerate the review of the filing, get contact information for the SME that ran the model and an SME from the vendor. The "SME" can be an intermediary at the insurer, e.g., a filing specialist, who can place the regulator in direct contact with the appropriate SMEs at the insurer or model vendor.	
A.2.e	~~If using catastrophe model output (a sub-model) as input to the GLM under review, verify whether loss associated with the modeled output was removed from the loss experience datasets.~~ Obtain an explanation of how catastrophic models are integrated into the model to ensure no double-counting.	1	If a weather-based sub-model is input to the GLM under review, loss data used to develop the model should not include loss experience associated with the weather-based sub-model. Doing so could cause distortions in the modeled results by double counting such losses when determining relativities or loss loads in the filed rating plan. For example, redundant losses in the data may occur when non-hurricane wind losses are included in the data while also using a severe convective storm model in the actuarial indication. Such redundancy may also occur with the inclusion of fluvial or pluvial flood losses when using a flood model, inclusion of freeze losses when using a winter storm model or including demand surge caused by any catastrophic event. Note that, the rating plan or indications underlying the rating plan, may provide special treatment of large losses and non-modeled large loss events. If such treatments exist, the company should	This note is not discussing catastrophe model outputs but rather a general statement about large losses and should be removed from the Sub-Model section and moved into the Data section or at the very least the weight of this observation should be lowered.

			provide an explanation how they were handled. These treatments need to be identified and the company/regulator needs to determine whether model data needs to be adjusted. For example, should large BI losses, in the case of personal automobile insurance, be capped or excluded, or should large non-catastrophe wind/hail claims in home insurance be excluded from the model's training, test and validation data?	
A.2.f	If using output of any scoring algorithms, obtain a list of the variables used to determine the score. ~~and provide the source of the data used to calculate the score.~~	1	~~Any sub-model should be reviewed in the same manner as the primary model that uses the sub-model's output as input.~~ Depending on the result of item A.2.b, the importance of this item may be decreased.	If the scoring algorithm is purchased, an insurer may not have the right to disclose this information. It would be much more efficient to instead point to filing of sub-models. Additionally, similar to the comment above, given the numerous sub-models common in the industry, extending the regulator's review to the underlying model, in the same breadth and depth as the insurer's filed model, seems impractical. Finally, there should be clarity as to the difference between a "scoring model" and a "sub-model."
3. Adjustments to Data				
A.3.a	Determine if premium, exposure, loss or expense data were adjusted (e.g., developed, trended, adjusted for catastrophe experience or capped) and, if so, how? ~~Do the adjustments vary for different segments of the data and, if so, identify the segments and how was the data adjusted?~~	2	The rating plan or indications underlying the rating plan may provide special treatment of large losses and non-modeled large loss events. If such treatments exist, the company should provide an explanation how they were handled. These treatments need to be identified and the company/regulator needs to determine whether model data needs to be adjusted. For example, should large bodily injury (BI) liability losses in the case of personal automobile insurance be excluded, or should large non-catastrophe wind/hail claims in home insurance be excluded from the model's training, test and validation data? Look for anomalies in the data that should be addressed. For example, is there an extreme loss event in the data? If other processes were used to load rates for specific loss events, how is the impact of those losses considered? Examples of losses that can contribute to anomalies in the data are large losses or flood, hurricane or severe convective storm losses for personal automobile comprehensive or home insurance.	The way that an insurer adjusts premium could be trade secret and APCIA is concerned about confidentiality. Further, insurers may be willing to describe the process at a high level, but getting into each transformation to each variable, is extensive.

A.3.b	Identify adjustments that were made to aggregated data, e.g., transformations, binning and/or categorizations. If any, identify the name of the characteristic/variable and obtain a description of the adjustment.	1		Previous sections will detail the adjustments made, so providing a comparison is not necessary.
A.3.c	Ask for aggregated data (~~one data set of pre-adjusted/scrubbed data and one data set of~~ post-adjusted/scrubbed data) that allows the regulator to focus on the univariate distributions and compare raw data to adjusted/binned/transformed/etc. data.	4	This is most relevant for variables that have been "scrubbed" or adjusted. Though most regulators may never ask for aggregated data and do not plan to rebuild any models, a regulator may ask for this aggregated data or subsets of it. It would be useful to the regulator if the percentage of exposures and premium for missing information from the model data by category were provided. This data can be displayed in either graphical or tabular formats.	If a regulator may never ask for the aggregated information, as pointed out in the note, what is the purpose of including it?
A.3.d	Determine how missing data was handled.	1	This is most relevant for variables that have been "scrubbed" or adjusted. The regulator should be aware of assumptions the modeler made in handling missing, null or "not available" values in the data. A statement on missing value treatment should be adequate. For example, it would be helpful to the reviewer if the modeler were to provide a statement as to whether there is any systemic reason for missing data. If adjustments or re-coding of values were made, they should be explained. It may also be useful to the regulator if the percentage of exposures and premium for missing information from the model data were provided. This data can be displayed in either graphical or tabular formats.	
A.3.e	If duplicate records exist, determine how they were handled.	1		
A.3.f	Determine if there were any material outliers identified and subsequently adjusted during the scrubbing process.	3	Look for a discussion of how outliers were handled. If necessary, the regulator may want to investigate further by getting a list (with description) of the outliers and determine what adjustments were made to each outlier. To understand the filer's response, the regulator should ask for the filer's materiality standard.	Drafters should make it clear that this section is looking for a description of the type of outliers and the treatment and not a listing.
4. Data Organization				

A.4.a	Obtain documentation on the methods used to compile and organize data, ~~including procedures to merge data from different sources or filter data based on particular characteristics~~ and a description of any preliminary analyses, data checks, and logical tests performed on the data and the results of those tests.	~~2~~ 3	This should explain how data from separate sources was merged or how subsets of policies, based on selected characteristics, are filtered to be included in the data underlying the model and the rationale for that filtering.	We are unclear as to the value of documenting the procedure to merge data as such the value of the information is disproportionate to the work needed to provide this information. At the very least this question should not be asked often and should be a level 3 or 4, instead of a 2.
A.4.b	~~Obtain documentation on the insurer's process for reviewing the appropriateness, reasonableness, consistency and comprehensiveness of the data, including a discussion of the rational relationship the data has to the predicted variable.~~	~~2~~	~~An example is when by peril or by coverage modeling is performed; the documentation should be for each peril/coverage and make rational sense. For example, if "murder" or "theft" data are used to predict the wind peril, provide support and a rational explanation for their use.~~	This question gets into the debate on correlation versus causation. Is actuarial justification sufficient or does this imply a logical argument is required? Insurers can provide information related to accuracy, consistency, and comprehensiveness, but have concerns with expanding the inquiry into causational questions. Further, the Information Element and Comment do not match. The information element is talking more about data as a whole, while the comment is going to specific variables, which is something better suited for "Building the Model." The comment does not reflect what the perceived intent of the information element is.
A.4.c	~~Identify material findings the company had during their data review and obtain an explanation of any potential material limitations, defects, bias or unresolved concerns found or believed to exist in the data. If issues or limitations in the data influenced modeling analysis and/or results, obtain a description of those concerns and an explanation how modeling analysis was adjusted and/or results were impacted.~~	~~1~~	~~A response of "none" or "n/a" may be an appropriate response.~~	What is the justification for this request? Is the request for the final elements or for the exploratory dataset? Bias in data is almost impossible to know without a full dataset to compare to – we can only make assumptions or inferences. The question does not necessarily speak to the appropriateness of the model. For example, where a new data element is only available on 40% of the exposure – how does this impact the model or output?

B. BUILDING THE MODEL

Section	Information Element	Level of Importance to Regulator's Review	Comments	APCIA comments
1. High-Level Narrative for Building the Model				
B.1.a	Identify the type of model underlying the rate filing (e.g. Generalized Linear Model – GLM, decision tree, Bayesian Generalized Linear Model, Gradient-Boosting Machine, neural network, etc.). Understand the model's role in the rating system and provide the reasons why that type of model is an appropriate choice for that role.	1	It is important to understand if the model in question is a GLM, and therefore these information elements are applicable or, if it is some other model type, in which case other reasonable review approaches may be considered. There should be an explanation of why the model (using the variables included in it) is appropriate for the line of business. If by-peril or by-coverage modeling is used, the explanation should be by-peril/coverage. Note, if the model is not a GLM, the information elements in this white paper may not apply in their entirety.	APCIA suggests that there should be additional clarity, if not deleted, as to the expectation for asking for an explanation of why the model is appropriate for the line of business. Different techniques may be suitable depending on the intended application of the model, however this is not specifically related to the line of business.
B.1.b	Identify the software used for model development. Obtain the name of the software vender/developer, software product ~~and a software version reference used in model development~~.	3	Changes in software from one model version to the next may explain if such changes, over time, contribute to changes in the modeled results. The company should provide the name of the third-party vendor and a "contact" in the event the regulator has questions. The "contact" can be an intermediary at the insurer who can place the regulator in direct contact with appropriate SMEs. Open-source software/programs used in model development should be identified by name ~~and version~~ the same as if from a vendor.	What is the purpose for this information? It is unclear how the software used for building the molde is relevant to the review and B.1.a already askes for information about the tpe of model. Without reviewing the code, which is impractical, it is unclear what the regulator will do with the version information. APCIA is uncertain as to how this applies to whether or not the model and output are appropriate. Also, if the insurer has developed their own internal tools, how will this be disclosed since those tools are proprietary. Ultimately, we feel this section should be deleted in its entirety. Otherwise, we offer a few suggested edits for your consideration.

B.1.c	Obtain a description how the available data was divided between model training, test and/or validation datasets. The description should include an explanation why the selected approach was deemed most appropriate, whether the company made any further subdivisions of available data and reasons for the subdivisions (e.g., a portion separated from training data to support testing of components during model building). Determine if the validation data was accessed before model training was completed and, if so, obtain an explanation why that came to occur. Obtain a discussion of whether the model was rebuilt using all of the data or if it was only based on the training data.	1	The reviewer should be aware that modelers may break their data into three or just two datasets. Although the term "training" is used with little ambiguity, "test" and "validation" are terms that are sometimes interchanged, or the word "validation" may not be used at all. It would be unexpected if validation and/or test data were used for any purpose other than validation and/or test, prior to the selection of the final model.	Cross validation is commonly used wen there is limited data available. Cross validation should be listed as an option in this section.
B.1.d	Obtain a brief description of the development process, from initial concept to final model and filed rating plan.	1	The narrative should have the same scope as the filing.	
B.1.e	Obtain a narrative on whether loss ratio, pure premium or frequency/severity analyses were performed and, if separate frequency/severity modeling was performed, how pure premiums were determined.	1		There should be an option for "deviation from current rate" in this list as well.
B.1.f	Identify the model's target variable.	1	A clear description of the target variable is key to understanding the purpose of the model. It may also prove useful to obtain a sample calculation of the target variable in Excel format, starting with the "raw" data for a policy, or a small sample of policies, depending on the complexity of the target variable calculation.	
B.1.g	Obtain a description of the variable selection process.	1	The narrative regarding the variable selection process may address matters such as the criteria upon which variables were selected or omitted, identification of the number of preliminary variables considered in developing the model versus the number of variables that remained, and any statutory or regulatory limitations that were taken into account when making the decisions regarding variable selection. ~~The modeler should comment if any form of data mining to identify selected variables was performed and explain how the modeler~~	It is unnecessary to comment on the data mining, because the general model validation should cover this.

			~~addressed "false positives" which often arise from data mining techniques.~~	
B.1.h	In conjunction with variable selection, obtain a narrative on how the company determine the granularity of the rating variables during model development.	3	This discussion should include discussion of how credibility was considered in the process of determining the level of granularity of the variables selected.	
B.1.i	~~Determine if model input data was segmented in any way. For example, was modeling performed on a by coverage, by peril, or by form basis? If so, obtain a description of data segmentation and the reasons for data segmentation.~~	~~1~~	~~The regulator would use this to follow the logic of the modeling process.~~	This information is duplicative of B.1.a
B.1.j	If adjustments to the model were made based on credibility considerations, obtain an explanation of the credibility considerations and how the adjustments were applied.	2	Adjustments may be needed given models do not explicitly consider the credibility of the input data or the model's resulting output; models take input data at face value and assume 100% credibility when producing modeled output.	The use of "credibility" seems to be referring to the actuarial concept of credibility. It is not clear what the relevance is of that concept in modeling. Rather, the question should focus on approach and philosophy to feature engineering in model statistics to determine the appropriateness of features and levels. The comments should clarify that this refers to model credibility, not actuarial. If this information element is trying to address credibility as it relates to adjustments made to model factors/results post model building, then this information may be more relevant in the filed rating plan section where selections are made to the factors due to credibility concerns.
2. Medium-Level Narrative for Building the Model				
B.2.a	At crucial points in model development, if selections were made among alternatives regarding model assumptions or techniques, obtain a narrative on the judgment used to make those selections.	3 4		This should be a level 4 instead of a 3. If a state has serious concerns with a model, this does not seem like an item that would come up.

B.2.b	~~If post-model adjustments were made to the data and the model was rerun, obtain an explanation on the details and the rationale for those adjustments.~~	~~2~~	~~Evaluate the addition or removal of variables and the model fitting. It is not necessary for the company to discuss each iteration of adding and subtracting variables, but the regulator should gain a general understanding how these adjustments were done, including any statistical improvement measures relied upon.~~	This is duplicative of B.1.d, so long as there is no clarity in when the "beginning" and "end" of a modeling exercise. APCIA does not believe this information is relevant if focused on the iterations of model building. As an example, is a model refresh considered a new modeling exercise or a continuation/update to an open modeling exercise?
B.2.c	Obtain a description of the testing that was performed during the model-building process ~~and a discussion of why interaction terms were included (or not included)~~.	3	There should be a description of testing that was performed during the model-building process. Examples of tests that may have been performed include univariate testing and review of a correlation matrix.	The interaction element is irrelevant unless a regulator has concerns around the interaction that was recommended to be included.
B.2.d	For the GLM, identify the link function used. Identify which distribution was used for the model (e.g., Poisson, Gaussian, log-normal, Tweedie). Obtain an explanation why the link function and distribution were chosen. ~~Obtain the formulas for the distribution and link functions, including specific numerical parameters of the distribution. Obtain a discussion of applicable convergence criterion.~~	1	Solving the GLM is iterative and the modeler can check to see if fit is improving. At some point convergence occurs, though when it occurs can be subjective or based on threshold criteria. The convergence criterion should be documented with a brief explanation of why it was selected. If the software's default convergence criteria were relied upon, the regulator should look for a description of the default convergence criterion and an explanation of any deviation from it.	~~The request for formulas for the distribution and link functions seems like unnecessary "textbook" information. Additionally, obtaining a description of convergence criterion does not provide any valuable information and should be deleted.~~
B.2.e	Obtain a narrative on the formula relationship between the data and the model outputs, with a definition of each model input and output. The narrative should include all coefficients necessary to evaluate the predicted pure premium, relativity or other value, for any real or hypothetical set of inputs.	2		
B.2.f	If there were data situations in which GLM weights were used, obtain an explanation of how and why they were used.	3	Investigate whether identical records were combined to build the model.	Are there level 1 or level 2 Information elements that could get to this information?
3. Predictor Variables				
B.3.a	Obtain ~~a complete data dictionary, including~~ the names, types, definitions ~~and uses~~ of each predictor variable, offset variable, control variable, proxy variable, geographic variable, geodemographic variable and all other variables in the model used on their own or as an interaction with other variables (including sub-models and external models)..	1	~~Types of variables might be continuous, discrete, Boolean, etc. Definitions should not use programming language or code. For any variable(s) intended to function as a control or offset, obtain an explanation of its purpose and impact. Also, for any use of interaction between variables, obtain an explanation of its rationale and impact.~~	Certainly, key variables must be clearly explained, but such a comprehensive and formal dictionary is not necessary. Including "uses" also raises confidentiality concerns.

B.3.b	~~Obtain a list of predictor variables considered but not used in the final model, and the rationale for their removal.~~	4	~~The purpose of this requirement is to identify variables that the company finds to be predictive but ultimately may reject for reasons other than loss cost considerations (e.g., price optimization). Also, look for variables the company tested and then rejected. This item could help address concerns about data dredging. The reasonableness of including a variable with given significance level could depend greatly on the other variables the company evaluated for inclusion in the model and the criteria for inclusion or omission. For instance, if the company tested 1,000 similar variables and selected the one with the lowest p-value of 0.001, this would be a far, far weaker case for statistical significance than if that variable was the only one the company evaluated. Note, context matters.~~	APCIA appreciates the intent behind this but there may be an extensive list of variables not considered, which could significantly detract from the primary focus of the regulator's review, which is to assess the model presented including the variables that were ultimately selected. Additionally, going through this list of variables that will have no impact on a customer's premium in the final rating plan is extensive and not relevant. The fact that an insurer looked at a variable that was later deemed to be rejected for some reason, should not have a bearing on the validity and approval of the model within the filing.
B.3.c	Obtain a correlation matrix for all predictor variables included in the model and sub-model(s).	3	While GLMs accommodate collinearity, the correlation matrix provides more information about the magnitude of correlation between variables. The company should indicate what statistic was used (e.g., Pearson, Cramer's V). The regulatory reviewer should understand what statistic was used to produce the matrix but should not prescribe the statistic.	
B.3.d	Obtain a rational explanation for why an increase in each predictor variable should increase or decrease frequency, severity, loss costs, expenses, or any element or characteristic being predicted.	3	The explanation should go beyond demonstrating correlation. Considering possible causation may be relevant, but proving causation is neither practical nor expected. If no rational explanation can be provided, greater scrutiny may be appropriate. For example, the regulator should look for unfamiliar predictor variables and, if found, the regulator should seek to understand the connection that variable has to increasing or decreasing the target variable.	This Information element is vague. Most of the rational explanations will relate to the correlation. Since causation cannot be proven it potentially can come with biases or misunderstanding of the characteristics. For example, the data can suggest as a predictor increases so does the loss cost and thus the predicted factors, however, to opine on potential reasonings for this may not be appropriate. We should let the data drive the discussion. However, a welcomed discussion on how the data aligns is appropriate. To ask if there is a frequency or severity component, how a variable interacts with those variables (for example, limit), or is there are large losses that drive the results are appropriate. This demonstrates that the modeler spent time to understand the data and what is driving the results that they see.

B.3.e	~~If the modeler made use of one or more dimensionality reduction techniques, such as a Principal Component Analysis (PCA), obtain a narrative about that process, an explanation why that technique was chosen, and a description of the step-by-step process used to transform observations (usually correlated) into a set of linearly uncorrelated variables. In each instance, obtain a list of the pre-transformation and post-transformation variable names, and an explanation how the results of the dimensionality reduction technique was used within the model.~~	~~2~~		This element is textbook "rules based" and redundant.
4. Adjusting Data, Model Validation and Goodness-of-Fit Measures				
B.4.a	Obtain a description of the methods used to assess the statistical significance/goodness of the fit of the model to validation data, such as lift charts and statistical tests. Compare the model's projected results to historical actual results and verify that modeled results are reasonably similar to actual results from validation data.	1	For models that are built using multi-state data, validation data for some segments of risk is likely to have low credibility in individual states. Nevertheless, some regulators require model validation on State-only data, especially when analysis using state-only data contradicts the countrywide results. State-only data might be more applicable but could also be impacted by low credibility for some segments of risk. ~~Look for geographic stability measures, e.g., across states or territories within state.~~	It is excessive to have the statement about geographic stability measures when the 1st paragraph talks about state level data being potentially low credibility.
B.4.b	~~For all variables (discrete or continuous), review the appropriate parameter values, confidence intervals, chi-square tests, p-values and any other relevant and material tests. Determine if model development data, validation data, test data or other data was used for these tests.~~	~~1~~	~~Typical p-values greater than 5% are large and should be questioned. Reasonable business judgment can sometimes provide legitimate support for high p-values. Reasonableness of the p-value threshold could also vary depending on the context of the model. For example, the threshold might be lower when many candidate variables were evaluated for inclusion in the model.~~ ~~Overall lift charts and/or statistical tests using validation data may not provide enough of the picture. If there is concern about one or more individual variables, the reviewer may obtain, for each discrete variable level, the parameter value, confidence intervals, chi-square tests, p-values and any other relevant and material tests. For variables that are modeled continuously, it may be sufficient to obtain statistics around~~	This is overly prescriptive and could include trade secret information. The information can be sufficiently gleaned elsewhere without this amount of detail.

			~~the modeled parameters; for example, confidence intervals around each level of an AOI curve might be more than what is needed.~~	
B.4.c	~~Identify the threshold for statistical significance and explain why it was selected. Obtain a reasonable and appropriately supported explanation for keeping the variable for each discrete variable level where the p-values were not less than the chosen threshold.~~	~~1~~	~~The explanation should clearly identify the thresholds for statistical significance used by the modeler. Typical p-values greater than 5% are large and should be questioned. Reasonable business judgment can sometimes provide legitimate support for high p-values. Reasonableness of the p-value threshold could also vary depending on the context of the model. For example, the threshold might be lower when many candidate variables were evaluated for inclusion in the model.~~ ~~Overall lift charts and/or statistical tests using validation data may not provide enough of the picture. If there is concern about one or more individual variables, the reviewer may obtain, for each discrete variable level, the parameter value, confidence intervals, chi-square tests, p-values and any other relevant and material tests.~~	This is overly prescriptive and could include trade secret information. The information can be sufficiently gleaned elsewhere without this amount of detail. Additionally, the information element is focused very specifically on p-values, yet in prior discussions it was mentioned that other tests may be leveraged as p-values may not be appropriate.

B.4.d	~~For overall discrete variables, review type 3 chi-square tests, p-values, F tests and any other relevant and material test. Determine if model development data, validation data, test data or other data was used for these tests.~~	~~2~~	~~Typical p-values greater than 5% are large and should be questioned. Reasonable business judgment can sometimes provide legitimate support for high p-values. Reasonableness of the p-value threshold could also vary depending on the context of the model, e.g., the threshold might be lower when many candidate variables were evaluated for inclusion in the model.~~ ~~Overall lift charts and/or statistical tests using validation data may not provide enough of the picture. If there is concern about one or more individual variables, the reviewer may obtain, for each discrete variable level, the parameter value, confidence intervals, chi-square tests, p-values and any other relevant and material tests. For variables that are modeled continuously, it may be sufficient to obtain statistics around the modeled parameters; for example, confidence intervals around each level of an AOI curve might be more than what is needed.~~	This is overly prescriptive and could include trade secret information. The information can be sufficiently gleaned elsewhere without this amount of detail.
B.4.e	~~Obtain evidence that the model fits the training data well, for individual variables, for any relevant combinations of variables and for, the overall model.~~	~~2~~	~~For a GLM, such evidence may be available using chi-square tests, p-values, F-tests and/or other means.~~ ~~The steps taken during modeling to achieve goodness-of-fit are likely to be numerous and laborious to describe, but they contribute much of what is generalized about GLM. We should not assume we know what they did and ask "how?" Instead, we should ask what they did and be prepared to ask follow-up questions.~~	This is overly prescriptive and could include trade secret information. The information can be sufficiently gleaned elsewhere without this amount of detail.

B.4.f	~~For continuous variables, provide confidence intervals, chi square tests, p-values and any other relevant and material test. Determine if model development data, validation data, test data or other data was used for these tests.~~	~~2~~	~~Typical p-values greater than 5% are large and should be questioned. Reasonable business judgment can sometimes provide legitimate support for high p-values. Reasonableness of the p-value threshold could also vary depending on the context of the model, e.g., the threshold might be lower when many candidate variables were evaluated for inclusion in the model.~~ ~~Overall lift charts and/or statistical tests using validation data may not provide enough of the picture. If there is concern about one or more individual variables, the reviewer may obtain, for each discrete variable level, the parameter value, confidence intervals, chi square tests, p-values and any other relevant and material tests. For variables that are modeled continuously, it may be sufficient to obtain statistics around the modeled parameters; for example, confidence intervals around each level of an AOI curve might be more than what is needed.~~	This is overly prescriptive and could include trade secret information. The information can be sufficiently gleaned elsewhere without this amount of detail.
B.4.g	Obtain a description how the model was tested for stability over time.	2	Evaluate the build/test/validation datasets for potential time-sensitive model distortions (e.g., a winter storm in year 3 of 5 can distort the model in both the testing and validation datasets). Obsolescence over time is a model risk (e.g., old data for a variable or a variable itself may no longer be relevant). If a model being introduced now is based on losses from years ago, the reviewer should be interested in knowing whether that model would be predictive in the proposed context. Validation using recent data from the proposed context might be requested. Obsolescence is a risk even for a new model based on recent and relevant loss data. The reviewer may want to inquire as to the following: What steps, if any, were taken during modeling to prevent or delay obsolescence? What controls will exist to measure the rate of obsolescence? What is the plan and timeline for updating and ultimately replacing the model? The reviewer should also consider that as newer technologies enter the market (e.g., personal automobile) their impact may change claim	

			activity over time (e.g., lower frequency of loss). So, it is not necessarily a bad thing that the results are not stable over time.	
~~B.4.h~~	~~Obtain a narrative on how potential concerns with overfitting were addressed.~~	~~2~~		This should be deleted as it is duplicative to B.4.a.
B.4.i	~~Obtain support demonstrating that the GLM assumptions are appropriate.~~	~~3~~	~~Visual review of plots of actual errors is usually sufficient.~~ ~~The reviewer should look for a conceptual narrative covering these topics: How does this particular GLM work? Why did the rate filer do what it did? Why employ this design instead of alternatives? Why choose this particular distribution function and this particular link function? A company response may be at a fairly high level and reference industry practices. If the reviewer determines that the model makes no assumptions that are considered to be unreasonable, the importance of this item may be reduced.~~	This is duplicative of B.2.d and is too theoretical/text book strict.
B.4.j	~~Obtain 5-10 sample records with corresponding output from the model for those records.~~	4		This is not necessary for the model and is potentially more appropriate to ask for rating examples, which would not be GLM/Predictive model questions.

5. "Old Model" Versus "New Model"				
B.5.a	Obtain an explanation why this model is an improvement to the current rating plan. If it replaces a previous model, find out why it is better than the one it is replacing; determine how the company reached that conclusion and identify metrics relied on in reaching that conclusion. Look for an explanation of any changes in calculations, assumptions, parameters, and data used to build this model from the previous model.	2	Regulators should expect to see improvement in the new class plan's predictive ability or other sufficient reason for the change.	
B.5.b	~~Determine if two Gini coefficients were compared and obtain a narrative on the conclusion drawn from this comparison.~~	~~3~~	~~One example of a comparison might be sufficient.~~ ~~This is relevant when one model is being updated or replaced. Regulators should expect to see improvement in the new class plan's predictive ability. This information element requests a comparison of Gini coefficient from the prior model to the Gini coefficient of proposed model. It is expected that there should be improvement in the Gini coefficient. A higher Gini coefficient indicates greater differentiation produced by the model and how well the model fits that data. This comparison is not applicable to initial model introduction. Reviewer can look to CAS monograph for information on Gini coefficients.~~	This is too prescriptive and should be eliminated.
B.5.c	~~Determine if double lift charts were analyzed and obtain a narrative on the conclusion drawn from this analysis.~~	~~2~~	~~One example of a comparison might be sufficient.~~ ~~Note that "not applicable" is an acceptable response.~~	This is to prescriptive and should be deleted.
B.5.d	If replacing an existing model, obtain a list of any predictor variables used in the old model that are not used in the new model. Obtain an explanation why these variables were dropped from the new model. Obtain a list of all new predictor variables in the new model that were not in the prior old model.	2	Useful to differentiate between old and new variables so the regulator can prioritize more time on variables not yet reviewed.	
6. Modeler Software				
B.6.a	Request access to SMEs (e.g., modelers) who led the project, compiled the data, and/or built the model. ~~and/or performed peer review.~~	~~3~~ 4	The filing should contain a contact that can put the regulator in touch with appropriate SMEs and key contributors to the model development to discuss the model.	This should be a level 4 as opposed to a 3. At a level 3 requesting access to the modelers appears excessive. Also requesting access to the SMEs that performed a peer review is quite excessive and

				would require companies to think about how to ensure the peer reviewer is prepared for discussions.

C. THE FILED RATING PLAN

Section	Information Element	Level of Importance to	Comments	
1. General Impact of Model on Rating Algorithm				
C.1.a	In the actuarial memorandum or explanatory memorandum, for each model and sub-model (including external models), look for a narrative that explains each model and its role (how it was used) in the rating system.	1	The "role of the model" relates to how the model integrates into the rating plan as a whole and where the effects of the model are manifested within the various components of the rating plan. This is not intended as an overarching statement of the model's goal, but rather a description of how specifically the model is used. This item is particularly important, if the role of the model cannot be immediately discerned by the reviewer from a quick review of the rate and/or rule pages. (Importance is dependent on state requirements and ease of identification by the first layer of review and escalation to the appropriate review staff.)	
C.1.b	~~Obtain an explanation of how the model was used to adjust the rating algorithm.~~	~~1~~	~~Models are often used to produce factor based indications, which are then used as the basis for the selected changes to the rating plan. It is the changes to the rating plan that create impacts. Consider asking for an explanation of how the model was used to adjust the rating algorithm.~~	This information is already addressed in section B.
C.1.c	Obtain a complete list of characteristics/variables used in the proposed rating plan, including those used as input to the model (including sub-models and composite variables) and all other characteristics/variables (not input to the model) used to calculate a premium. For each characteristic/variable, determine if it is only input to the model, whether it is only a separate univariate rating characteristic, or whether it is both input to the model and a separate univariate rating characteristic. The list should include transparent descriptions (in plain	1	Examples of variables used as inputs to the model and used as separate univariate rating characteristics might be criteria used to determine a rating tier or household composite characteristic.	

	language) of each listed characteristic/variable.			
2. Relevance of Variables and Relationship to Risk of Loss				
C.2.a	Obtain a narrative regarding how the characteristics/rating variables included in the filed rating plan relate to the risk of insurance loss (or expense) for the type of insurance product being priced.	2	~~The narrative should include a discussion of the relevance each characteristic/rating variable has on consumer behavior that would lead to a difference in risk of loss (or expense).~~ The narrative should include a rational relationship to cost, and model results should be consistent with the expected direction of the relationship. This explanation would not be needed if the connection between variables and risk of loss (or expense) has already been illustrated.	The narrative is going to be subjective, instead the data should just speak for itself.
3. Comparison of Model Outputs to Current and Selected Rating Factors				
C.3.a	Compare relativities indicated by the model to both current relativities and the insurer's selected relativities for each risk characteristic/variable in the rating plan.	1	"Significant difference" may vary based on the risk characteristic/variable and context. However, the movement of a selected relativity should be in the direction of the indicated relativity; if not, an explanation is necessary as to why the movement is logical.	Insurers are willing to provide a high-level explanation of the selection, but not the factors due to confidentiality concerns. This Information Element may be more relevant on model refresh/updates to indications.
C.3.b	Obtain documentation and support for all calculations, judgments, or adjustments that connect the model's indicated values to the selected values.	1	The documentation should include explanations for the necessity of any such adjustments and explain each significant difference between the model's indicated values and the selected values. This applies even to models that produce scores, tiers, or ranges of values for which indications can be derived. This information is especially important if differences between model indicated values and selected values are material and/or impact one consumer population more than another.	This Information Element is duplicative of C.3.a.

C.3.c	For each characteristic/variable used as both input to the model (including sub-models and composite variables) and as a separate univariate rating characteristic, obtain a narrative how each characteristic/variable was tempered or adjusted to account for possible overlap or redundancy in what the characteristic/variable measures.	2	Modeling loss ratio with these characteristics/ variables as control variables would account for possible overlap. The insurer should address this possibility or other considerations, e.g., tier placement models often use risk characteristics/ variables that are also used elsewhere in the rating plan. One way to do this would be to model the loss ratios resulting from a process that already uses univariate rating variables. Then the model/composite variables would be attempting to explain the residuals.	
4. Responses to Data, Credibility and Granularity Issues				
C.4.a	Determine what, if any, consideration was given to the credibility of the output data.	2	At what level of granularity is credibility applied. If modeling was by-coverage, by-form or by-peril, explain how these were handled when there was not enough credible data by coverage, form or peril to model.	
C.4.b	If the rating plan is less granular than the model, obtain an explanation why.	2	This is applicable if the insurer had to combine modeled output in order to reduce the granularity of the rating plan.	Consider combining this C.4.b and C.4.c.
C.4.c	If the rating plan is more granular than the model, obtain an explanation why.	2	A more granular rating plan may imply that the insurer had to extrapolate certain rating treatments, especially at the tails of a distribution of attributes, in a manner not specified by the model indications. However, it may be necessary to extrapolate due to data availability or other considerations.	APCIA believes the comment for this information element is presumptive and needs to be balanced out with recognition that there could be other explanations.
5. Definitions of Rating Variables				
C.5.a	~~Obtain a narrative on adjustments made to model output, e.g., transformations, binning and/or categorizations. If adjustments were made, obtain the name of the characteristic/variable and a description of the adjustment.~~	~~2~~	~~If rating tiers or other intermediate rating categories are created from model output, the rate and/or rule pages should present these rating tiers or categories. The company should provide an explanation how model output was translated into these rating tiers or intermediate rating categories.~~	This is too detailed and it is not clear how this provides value to the review. It also could be trade secret. There is a risk of exposing intellectual property without serving any benefit. This is simply more documentation and does not help in the review.
6. Supporting Data				
C.6.a	~~Obtain aggregated state-specific, book of business-specific univariate historical experience data, separately for each year included in the model, consisting of loss ratio or pure premium relativities and the data underlying those calculations for each category of model output(s) proposed to be used within the rating plan. For each data element, obtain an explanation whether it is raw or adjusted and, if the latter, obtain a detailed~~	4	~~For example, were losses developed/undeveloped, trended/untrended, capped/uncapped, etc.?~~ ~~Univariate indications should not necessarily be used to override more sophisticated multivariate indications. However, they do provide additional context and may serve as a useful reference.~~	Caution must be exercised whenever drilling down on actual results, including by year, as these may volatile and in and of themselves not reflective of the model's predictive power. This variability should be visible and contemplated through appropriate statistical metrics and testing. Additionally, there are confidentiality considerations. While this does not have individual record consumer information, with enough of these data cuts a competitor could back into

	~~explanation for the adjustments.~~			the model and rating plan much easier.
C.6.b	Obtain an explanation of any material (especially directional) differences between model indications and state-specific univariate indications.	4	Multivariate indications may be reasonable as refinements to univariate indications, but possibly not for bringing about significant reversals of those indications. For instance, if the univariate indicated relativity for an attribute is 1.5 and the multivariate indicated relativity is 1.25, this is potentially a plausible application of the multivariate techniques. If, however, the univariate indicated relativity is 0.7 and the multivariate indicated relativity is 1.25, a regulator may question whether the attribute in question is negatively correlated with other determinants of risk. Credibility of state data should be considered when state indications differ from modeled results based on a broader data set. However, the relevance of the broader data set to the risks being priced should also be considered. Borderline reversals are not of as much concern.	Consider adding a comment that if the multivariate performs well against the state level data, then this should suffice. However, credibility considerations need to be made as state level segmentation comparisons generally do not have enough credibility.
7. Consumer Impacts – This is an important consideration for implementing a plan. However, this is not a statistical concept and reliance upon this may actually lead to rates that are not cost based. Rather than an obstacle to approval we suggest consumer impacts must be clearly understood, and the regulator and the company must work together to develop an implementation plan that addresses any concerns. However, these should not be the basis for evaluating the predictive model itself. Additionally, there should be more clarity as to what is meant by "consumer."				
C.7.a	~~Obtain a listing of the top five rating variables that contribute the most to large swings in premium, both as increases and decreases.~~	~~4~~	~~These rating variables may represent changes to rating factors, be newly introduced to the rating plan, or have been removed from the rating plan.~~	This is complicated to answer. A company would need additional objective information before they could answer this. For example, does a large swing refer to a renewal or difference between two customers?
~~C.7.b~~	~~Determine if the insurer performed sensitivity testing to identify significant changes in premium due to small or incremental change in a single risk characteristic. If such testing was performed, obtain a narrative that discusses the testing and provides the results of that testing.~~	~~3~~	~~One way to see sensitivity is to analyze a graph of each risk characteristic's/variable's possible relativities. Look for significant variation between adjacent relativities and evaluate if such variation is reasonable and credible.~~	This is also complicated to answer and not all transition testing can be done and requires a lot of simulations. Not all companies will have this capability.
C.7.c	For the proposed filing, obtain the impacts on expiring policies and describe the process used by management, if any, to mitigate those impacts.	2	Some mitigation efforts may substantially weaken the connection between premium and expected loss and expense, and hence may be viewed as unfairly discriminatory by some states.	

C.7.d	Obtain a rate disruption/dislocation analysis, demonstrating the distribution of percentage and/or dollar impacts on renewal business (created by rerating the current book of business), and sufficient information to explain the disruptions to individual consumers.	2	The analysis should include the largest dollar and percentage impacts arising from the filing, including the impacts arising specifically from the adoption of the model or changes to the model as they translate into the proposed rating plan. While the default request would typically be for the distribution/dislocation of impacts at the overall filing level, the regulator may need to delve into the more granular variable-specific effects of rate changes if there is concern about particular variables having extreme or disproportionate impacts, or significant impacts that have otherwise yet to be substantiated. See Appendix C for an example of a disruption analysis.	
C.7.e	Obtain exposure distributions for the model's output variables and show the effects of rate changes at granular and summary levels, including the overall impact on the book of business.	3	See Appendix C for an example of an exposure distribution.	Item C.7.d should suffice for this. The impacts shown here are combined with all changes occurring, not just those related to variables in the model.
C.7.f	Identify policy characteristics, used as input to a model or sub-model, that remain "static" over a policy's lifetime versus those that will be updated periodically. Obtain a narrative on how the company handles policy characteristics that are listed as "static," yet change over time.	3	Some examples of "static" policy characteristics are prior carrier tenure, prior carrier type, prior liability limits, claim history over past X years, or lapse of coverage. These are specific policy characteristics usually set at the time new business is written, used to create an insurance score or to place the business in a rating/underwriting tier, and often fixed for the life of the policy. The reviewer should be aware, and possibly concerned, how the company treats an insured over time when the insured's risk profile based on "static" variables changes over time but the rate charged, based on a new business insurance score or tier assignment, no longer reflect the insured's true and current risk profile. A few examples of "non-static" policy characteristics are age of driver, driving record and credit information (FCRA related). These are updated automatically by the company on a periodic basis, usually at renewal, with or without the policyholder explicitly informing the company.	

C.7.g	~~Obtain a means to calculate the rate charged a consumer.~~	~~3~~	~~The filed rating plan should contain enough information for a regulator to be able to validate policy premium. However, for a complex model or rating plan, a score or premium calculator via Excel or similar means would be ideal, but this could be elicited on a case-by-case basis. Ability to calculate the rate charged could allow the regulator to perform sensitivity testing when there are small changes to a risk characteristic/variable. Note that this information may be proprietary.~~	This is logistically challenging for companies to execute and it is more a market conduct item than a filing review item. Additionally, the rate order of calculation rule around ration policy should suffice here.
C.7.h	~~In the filed rating plan, be aware of any non-insurance data used as input to the model (customer provided or other). In order to respond to consumer inquiries, it may be necessary to inquire as to how consumers can verify their data and correct errors.~~	~~1~~	~~If the data is from a third party source, the company should provide information on the source. Depending on the nature of the data, data may need to be documented with an overview of who owns it and the topic of consumer verification may need to be addressed, including how consumers can verify their data and correct errors.~~	If this information element is concerning the non-insurance (3rd party data) that is used in the development of the model, then other Section A would already have discussed this item. If this information element is concerning the non-insurance (3rd party data) that is used to rate a policy, then we would request replacing "as input to the model" with "as variables in the rating of a policy" for clarity.
8. Accurate Translation of Model into a Rating Plan				
C.8.a	~~Obtain sufficient information to understand how the model outputs are used within the rating system and to verify that the rating plan's manual, in fact, reflects the model output and any adjustments made to the model output.~~	~~1~~	~~The regulator can review the rating plan's manual to see that modeled output is properly reflected in the manual's rules, rates, factors, etc.~~	Is this information element asking for a new piece of documentation or is it just recommending that the regulator should seek to understand this information?
9. Efficient and Effective Review of Rate Filing				
C.9.a	Establish procedures to efficiently review rate filings and models contained therein.	1	"Speed to market" is an important competitive concept for insurers. Though regulators need to understand the rate filing before accepting the rate filing, the regulator should not request information which does not increase their understanding of the rate filing. Regulators should review their state's rate filing review process and procedures to ensure that they are fair and efficient. Regulators need to be aware that requesting information that is not necessary for a decision to be made on a rate filing's compliance with state laws and regulations.	The last sentence in the second paragraph of the comment is incomplete.
C.9.b	Be knowledgeable of state laws and regulations in order to determine if the proposed rating plan (and models) are compliant with state law.	1	This is a primary duty of regulators. The regulator should be knowledgeable of their state laws and regulations and apply them to a rate filing fairly and efficiently. The regulator should pay special attention to prohibitions of unfair discrimination.	

C.9.c	Be knowledgeable of state laws and regulations in order to determine if any information contained in the rate filing (and models) should be treated as confidential.	1	The regulator should be knowledgeable of their state laws and regulations regarding confidentiality of rate filing information and apply them to a rate filing fairly and efficiently. Confidentiality of proprietary information is key to innovation and competitive markets.	

491 22nd Ave. N.
St. Petersburg, FL 33704
908.265.5272

Free markets. Real solutions.
www.rstreet.org

July 27, 2020

Kris DeFrain
NAIC Casualty Actuarial and Statistical Task Force
Filed Electronically Via: kdefrain@naic.org

RE: CAS Task Force Draft White Paper, Exposed 6-12-2020

Dear Ms. DeFrain,

I write you as director of finance, insurance and trade policy at the R Street Institute, a nonprofit, nonpartisan public policy research organization ("think tank"). We appreciate the opportunity afforded by the Casualty Actuarial and Statistical Task Force to offer input on the revised draft of the Regulatory Review of Predictive Models White Paper.

We commend the Task Force for its work and are heartened by several updates in the revised paper. For example, in the section on relevance of variables and relationship to risk of loss, we welcome the revised paper's substitution that a rate-filing narrative ought to explain a variable's "rational" relationship to cost, rather than original "logical and intuitive." Actuarially credible and statistically significant variables may, in fact, prove to be counter-intuitive.

However, we remain concerned that the paper exceeds the scope of its stated purpose and that it could be interpreted to recommend more stringent reviews of existing models that have served consumers well for decades. In response to earlier comments we filed, the Ad Hoc Team asserted "the fact that predictive models have been reviewed in depth by regulators for many years under the current confidentiality provisions…is prima facie evidence that the negative impacts that R Street is alleging will not arise." However, the team elsewhere notes that it "would be unfortunate if a model is withdrawn from the market because the modeler is unwilling to share information with regulators."

Our position is that it would be not just unfortunate, but disruptive, and threatens to reverse progress made over three decades toward more competitive insurance markets that better serve consumers. We will continue to recommend that regulators exercise to avoid such market disruption.

Sincerely,

R.J. Lehmann
Director of Finance, Insurance and Trade Policy
R Street Institute

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\CAS Task Force Comments

Consumer Data Industry Association
1090 Vermont Ave., NW, Suite 200
Washington, D.C. 20005-4905

P 202 371 0910

Writers email:
eellman@cdiaonline.org Writer's
direct dial: +1 (202) 408-7407
CDIAONLINE. ORG

Honorable Steve Kelley
Commissioner, Minnesota Department of Commerce
Chairman, NAIC Casualty Actuarial and Statistical Task Force
Minnesota Department of Commerce
85 7th Place East, Suite 280
Saint Paul, MN 55101

Honorable James J. Donelon
Commissioner, Louisiana Department of Insurance
Vice-Chairman, NAIC Casualty Actuarial and Statistical Task Force
1702 N. Third Street; P.O. Box 94214
Baton Rouge, LA 70802

Submitted Electronically to kdefrain@naic.org

Re: Best Practices for Regulatory Review of Predictive Analytics White Paper
Dear Chairman Kelley and Vice Chair Donelon:

I write on behalf of the Consumer Data Industry Association (CDIA) to comment on the exposure draft concerning best practices when reviewing predictive models and analytics. This draft was released by your Casualty Actuarial and Statistical Task Force ("Task Force") on June 12, 2020. Thank you for allowing CDIA another chance to offer comments on behalf of our consumer reporting agency ("CRA") members. We offer comments on section VI in the body of the whitepaper and sections A, B and C in the modeling guide.

The Consumer Data Industry Association is the voice of the consumer reporting industry, representing consumer reporting agencies including the nationwide credit bureaus, regional and specialized credit bureaus, background check and residential screening companies, and others. Founded in 1906, CDIA promotes the responsible use of consumer data to help consumers achieve their financial goals, and to help businesses, governments and volunteer organizations avoid fraud and manage risk. Through data and analytics, CDIA members empower economic opportunity all over the world, helping ensure fair and safe transactions for consumers, facilitating competition and expanding consumers' access to financial and other products suited to their unique needs.

Section VI, 1. c (p. 5) addresses a "Review [of] the individual input characteristics to and output factors from the predictive model (and its sub-models), as well as, associated selected relativities to ensure they are not unfairly discriminatory". We appreciate your feedback on our initial comments expressing concerns related to including "sub-models" like Credit-Based Insurance Scores ("CBIS") into the regulatory

review process. However, we do respectfully believe this will increase the burden of regulatory compliance for CRAs, slowdown the speed to market and impede the relationship between insurers and consumers. These new burdens can inject unnecessary friction into consumers who seek quick decisions and competitive prices from their insurers.

We respectfully believe these are "new, proposed obligations". The review of CBIS models has been established and ongoing in many States for close to two decades like you highlight, but those occur in other forms of insurance and not under the forms the Casualty Actuarial and Statistical (C) Task Force is seeking to add to its handbook and make an industry wide practice. The current reviews may include the same CBIS models, but if they are not currently being reviewed then we would argue these are in fact new obligations on CRAs.

Many States have provided certain confidentiality protections from the general public for CBIS models in accordance with their State law, but many is not all states. CDIA members spend significant amounts of time and resources developing their models and complying with current regulations. only takes one employee in one state to make one mistake and decades of hard work, investment and research is available for anyone to view, replicate, deceive or use to commit fraud. We are encouraged by the inclusion of new confidentiality language in Section VII of the Whitepaper, pertaining state confidential, proprietary, or trade secret state laws and relevant contractual provisions, and request inclusion of the language as a proposed change to the Product Filing Review Handbook. Even with the new language, the lack of a national exemption from public records remains a concern because information that has never previously been requested could be subject to the myriad of public disclosure laws around the country. There is no surety to how all states will respond to public records requests.

New language in Section V of the Whitepaper suggests that reliance on state confidentiality authority, regulations, and rules may not govern if the NAIC or another third party becomes involved in the review process on behalf of the states. NAIC or third-party participation in the review process causes significant trade secret and proprietary information protection concerns. It is not clear from the new language what protections, law, or authority would apply in such a case. We request clarifying language be added that, as a floor, the confidential, proprietary, and trade secret protections of the state on behalf of which a review is being performed apply.

We understand no information should be confidential from the regulators themselves. However, if the CBIS models are reviewed and accepted elsewhere, it would seem that a repetitive and costly process is occurring for not much if any added value to the final product for the consumers. The credit reporting system is a consistent nationwide process. Exposing individual characteristics of scoring models to public record requests allows competitors access to information that they can use to gain an unfair advantage over another company. It also reduces the incentive to continue to

create new solutions, reducing a competitive environment, which ultimately hurts consumers. Regulators should be able to know whether scoring models are in compliance with the law, but this information should not be accessible as a public record.

The potential for confidentiality concerns is not only with the CRAs, but the companies they work with (date furnishers and lenders) in the credit reporting system and their consumers. We are not convinced that including CBIS in this type of review is mission critical. Yet, if this review needs to be in the process, CDIA recommends the establishment of highly specific rules to protect confidentiality and proprietary information. Additionally, a separate review process of sub-models as an optional request with defined valid concerns would help in addressing concerns.

Credit-based insurance scores do not unfairly discriminatory towards any race, religion, gender, ethnicity, or other established suspect classes and there are studies that show the lack of illegal discrimination. A myth of illegal discrimination pervades many media accounts and public policy debates, but in truth, credit-based insurance scores do not promote redlining or other illegal insurance practices.

Section VI 3.a. (p. 6) addresses how to "[e]valuate how the model interacts with and improves the rating plan" and how to "[o]obtain a clear understanding of the characteristics that are input to a predictive model (and its sub-models), their relationship to each other and their relationship to non-modeled characteristics/variables used to calculate a risk's premium." We recognize the goal of the regulator in seeking to understand how the individual components of the rating plan interrelate to produce a consumer's premium, but we feel the NAIC's comment to CDIA comments on this add further confusion to our members. The white paper only mentions "characteristics", but your comment refers to "information that the 'CRAs use to create CBIS' is essential to understanding the structure of the CBIS models, the variables used, and their justification.". CRAs could provide general characteristics of the model without having confidentiality concerns, but the "information they use to create CBIS" appears to be far more specific.

If these provisions are meant to include information relating to the scoring models that CRAs use to create CBIS, there would be a significant new regulatory burden on CRAs and this would impede the relationship between insurers and consumers. These new burdensome requirements can inject unnecessary friction on to consumers who seek quick decisions and competitive prices from their insurers. Along with heightening the risk of disclosing proprietary information that is currently kept confidential because of its importance.

In "Selecting Model Input" under subsections A.1.a "Available Data Sources",

we appreciate the edits made to address our concerns around FCRA requirements being extended to all external data sources and the review for CBIS models being restricted to credit variables used in the model and not all credit variables.

We appreciate the task force's comment related to being open to changing the level for A.2.b"Determine if the sub-model was previously approved (or accepted) by the regulatory agency," the review level change is appreciated as it will eliminate unnecessary and duplicative reviews of third-party and vendor models that have been previously approved. To be consistent with the A.2.b review level change, a change from a review level 1 to a 3 or 4 is requested for current A.2.f, former A.2.e, "If using output of any scoring algorithms, obtain a list of the variables used to determine the score and provide the source of the data used to calculate the score".

Section A.4.c addresses "Identif[ing] material findings the company had during their data review and obtain an explanation of any potential material limitations, defects, bias or unresolved concerns found or believed to exist in the data. If issues or limitations in the data influenced modeling analysis and/or results, obtain a description of those concerns and an explanation how modeling analysis was adjusted and/or results were impacted". This provision should be recategorized from its current score of 1 to a 3 or 4 score. Existing regulations around actuarial rate making standards and state regulations should prevent these items from entering a "final/proposed" model. This should be categorized as three of four (i.e. if model review uncovers issues).

We have several comments regarding Section B, "building the model" :

- Sec. B.2.c, "Obtain a description of univariate balancing and the testing that was performed during the model-building process, including an explanation of the thought processes involved and a discussion of why interaction terms were included (or not included)."Only included interactions should be discussed. Interactions not be included, but default are not in a model, and therefore should not need to be justified.
- Secs. B.3.a and B.3.c., Both subsections pose trade secret protection and confidentiality issues.
- Sec. B.3.b, "Obtain[ing] a list of predictor variables considered but not used in the final model, and the rationale for their removal". The best practices and guidelines should be limited to only the variables that were in the final and proposed models.
- Sec. B.3.d, "Obtain[ing] an rational explanation for why an increase in each predictor variable should increase or decrease frequency, severity, loss costs, expenses, or any element or characteristic being predicted." CDIA agrees with the current and actuarially accepted practice of rate making guidelines not requiring intuitive or rational explanations of predictive values. We support use of variables that are statistically and actuarially predictive of insurance losses.

Additionally, this subsection poses a risk exposing trade secret and confidential information.

- Secs. B.4.b, through B.4.b CDIA recommends recategorizing these scores from their current scores of two to a three or four score, along with only making this a requirement if deemed necessary.
- Sec. B.4.c "Identif[ing] the threshold for statistical significance and explain why it was selected. Obtain a reasonable and appropriately supported explanation for keeping the variable for each discrete variable level where the p-values were not less than the chosen threshold". We thank NAIC for accepting our recommended language change of adding "threshold for statistical significance" into the list of required elements and changing this score from its current one to a three or four.

We have several comments regarding "Section C, "The Filed Rating Plan":

- Sec. C.1.c, like many other areas, this provision creates potential trade secret and confidentiality issues.
- Sec. C.7.h, we thank NAIC for easing the FCRA requirement section here.

The "Supporting Data" section, specifically Secs. C.6.a and C.6.b, on "Obtain[ing] an explanation of any material (especially directional) differences between model indications and state-specific univariate indications" pose some concerns for CRAs and could interfere with the insurance process for consumers.

Section VIII of the Whitepaper proposes several changes to the Handbook. Section X, "Other Considerations" of the Handbook suggest advisory organization regulation of model and algorithm vendors. As explained further in this comment, CIBS modelers are already heavily regulated.

Credit Based Insurance Scores are constructed using nationwide data sets. Scoring or grading their performance out at a state level may not be supported accurately with this approach. It is also a common occurrence for certain contracts to prevent model providers from sharing distinct or customer specific data with third parties. There are several factors besides credit information and CBIS that go into the rate setting process. Credit Information and CBIS may be the only ones that are consistent and transferrable across the country, while some of the other factors used can and do differ greatly on a state by state basis.

The insurance industry has been using CBIS models for decades and they have been approved by nearly every state's insurance department for auto and home insurers. Adding the work CASTF proposes will be burdensome and repetitive. The lack of trade secret and proprietary information protection will always remain a source of concern. In the long run we see this as something only large insurers will be able to absorb and the small to medium sized insurers that rely on third parties help will get squeezed out. We strongly feel that this will give large insurers a competitive edge in the marketplace. This will come at great cost to the consumers when their options decrease because of the eventual lack of competition.

There is already a large regulatory review presence on the industry. It is already over seen at the federal level by the Consumer Financial Protection Bureau (CFPB) and Federal Trade Commission (FTC), along with several states implementing their own regulations and the Conference of State Banking Commissioners looking into the industry as well. This increased regulation not only hurts the industry, but the consumers it serves. It will significantly hamper speed to market for the products consumers need and does not appear to add much, if any, benefit to the outcome for the industry and its consumer.

In conclusion, we believe that these potential new best practices will create burdensome regulatory difficulties for our members, speed to market issues for insurance companies, their product and the consumers that need them. CDIA members provide quality products that are already regulated and accepted by the insurance industry. CDIA and its members respectfully request consideration and inclusion of its comments in the task force's whitepaper. Thank you for the opportunity to comment and please feel free to contact us with any questions you may have.
Sincerely,

Eric J. Ellman
Senior Vice President, Public Policy & Legal Affairs

cc: Members of the Casualty Actuarial and Statistical Task Force (CASTF) of the Property and Casualty Insurance (C) Committee
Kris DeFrain, NAIC Staff
Jennifer Gardner, NAIC Staff

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\CDIA Comments

Comments for the Center for Economic Justice

To the Casualty Actuarial Task Force

Regulatory Review of Predictive Models White Paper

July 27, 2020

The Center for Economic Justice offers the following comments on the June 12, 2020 exposure draft of the "Regulatory Review of Predictive Models White Paper."

CEJ greatly appreciates the effort CASTF has expended to grapple with regulatory review of complex models. However, the most recent exposure misses the mark in at least two foundational ways.

1. The use of "rational explanation" is used incorrectly and inappropriately; and

2. The absence of guidance to address proxy discrimination against protected classes is a huge hole in regulatory review and a baffling omission.

1. Rational Explanation

The terms "rational explanation," "rational relationship" and "rational sense" are used several times in the exposure draft.

On page 12, as a topic for future discussion, the paper states, "Provide guidance for regulators that seek a causal or **rational explanation** why a rating variable is correlated to expected loss or expense, and why that correlation is consistent with the expected direction of the relationship."

On page 20 in Appendix A, "Data Organization" Section A.4.b, the paper states: "Obtain documentation on the insurer's process for reviewing the appropriateness, reasonableness, consistency and comprehensiveness of the data, including a discussion of the **rational relationship** the data has to the predicted variable," and "An example is when by-peril or by-coverage modeling is performed; the documentation should be for each peril/coverage and make **rational sense**. For example, if "murder" or "theft" data are used to predict the wind peril, provide support and a **rational explanation** for their use."

On page 34, in Appendix B, "Predictor Variables" Section B.3.d, the paper states, "Obtain a **rational explanation** for why an increase in each predictor variable should increase or decrease frequency, severity, loss costs, expenses, or any element or characteristic being predicted," and "The explanation should go beyond demonstrating correlation. Considering possible causation may be relevant, but proving causation is neither practical nor expected. If no **rational explanation** can be provided, greater scrutiny may be appropriate. For example, the regulator should look for unfamiliar predictor variables and, if found, the regulator should seek to understand the connection that variable has to increasing or decreasing the target variable."

On page 30,, in Appendix C, Section C 2 – Relevance of Variables and Relationship to Risk of Loss," the paper states, "Obtain a narrative regarding how the characteristics/rating variables included in the filed rating plan relate to the risk of insurance loss (or expense) for the type of insurance product being priced," and "The narrative should include a discussion of the relevance each characteristic/rating variable has on consumer behavior that would lead to a difference in risk of loss (or expense). The narrative should include a **rational relationship** to cost, and model results should be consistent with the expected direction of the relationship. This explanation would not be needed if the connection between variables and risk of loss (or expense) has already been illustrated."

On page 46, the paper defines "**Rational Explanation**" -- A "**rational explanation**" refers to a plausible narrative connecting the variable and/or treatment in question with real-world circumstances or behaviors that contribute to the risk of insurance loss in a manner that is readily understandable to a consumer or other educated layperson. A "**rational explanation**" does not require strict proof of causality but should establish a sufficient degree of confidence that the variable and/or treatment selected are not obscure, irrelevant, or arbitrary. A "**rational explanation**" can assist the regulator in explaining an approved rating treatment if challenged by a consumer, legislator, or the media. Furthermore, a "rational explanation" can increase the regulator's confidence that a statistical correlation identified by the insurer is not spurious, temporary, or limited to the specific data sets analyzed by the insurer."

It is important to note that while "**rational explanation"** is defined, the undefined terms "rational relationship and "rational sense" are also used.

The paper also uses the term "**rationale**" in several instances. "Rationale" is used in Section A.4.a, B.2.b., B.3.a and B3.b. In each instance, the term "**rationale"** is synonymous with "**explanation"** or "**justification**."

1.1 The use of "rational explanation" is inappropriate because it introduces subjective interpretation by the regulator in place of a valid statistical analysis to evaluate the potential for a spurious correlation.

The clear intent of the paper's use of "rational explanation" or similar terms is two-fold. First, it is used to mean the insurer must provide a "rationale" for the purported relationship. Second, it is used to mean that this "rationale" must be "reasonable" or "plausible" to the regulator.

The clear intent of the use of "rational explanation" is to identify spurious correlations – correlations[1] – "two or more events or variables that are associated but not causally related due to either coincidence or a third unseen factor."[2] The problem with a spurious correlation is that the purported relationship is either transitory or illusory. As discussed below, regulators have always had the authority and responsibility to identify and stop the use of classifications that are spuriously correlated to particular insurance outcomes. And as discussed below, spurious correlation can reflect and perpetuate systemic racism and lead to proxy discrimination against protected classes.

The problem with the use of "rational explanation" is that it is inherently subjective and, consequently, arbitrary. To point out the obvious, each of our views about what constitutes a reasonable or rational explanation is based in very large part on our cultural biases – where and how we grew up and what our life experiences have been. The Black Lives Matters movement – and the response of insurer CEO and NAIC leadership to address systemic racism and inherent bias and to start a dialogue on race and diversity – speaks to the fact that what might be seen as rational to one person is understood as irrational and racist by another. For example, leading up to the Civil War, many in the country viewed slavery as rational. In the 1950's and 1960s, many viewed segregation of races as rational.

These historical examples are, of course, relevant. But, we can look at insurance examples today. Over the past few weeks and months, insurers have been offering "rational explanations" why they shouldn't provide auto insurance premium relief or why the premium relief offered wasn't greater. In some states, regulators have simply accepted those explanations as "rational," while regulators in other states have found the rationales offered failed the test of reasonableness.

We fully support regulators' interest in and actions to identify and eliminate spurious correlations in complex, predictive models. But relying upon a "rational explanation" is simply not the way to go because it is inherently subjective and lacks scientific rigor.

[1] https://tylervigen.com/spurious-correlations
[2] https://en.wikipedia.org/wiki/Spurious_relationship

In place of "rational explanation," we suggest that the white paper utilize "rationale" when the intent is for the insurer to explain or justify a practice. We also suggest that the white paper include specific guidance for scientific inquiry not just when the regulator suspects a spurious correlation, but guidance for insurers to test for spurious correlations – particularly when the spurious correlation may be a proxy for discrimination against a protected class.

1.2 The NCOIL and industry trades' argument that "rational explanation" usurps legislative authority is without merit or evidence.

We are aware of NCOIL passing a resolution declaring that legislative intent has always been that, other than specific prohibitions of particular risk classifications, the only requirement for use of a risk classification is a correlation. Attached, please find CEJ's comments to NCOIL, in which we demonstrate that

- There is no support for this proposition is statutory or regulatory language
- There is no support for this proposition in actuarial standards of practice
- The statutory language in NAIC models clearly requires more than a simple correlation as justification for a risk classification
- Regulatory practice refutes the "correlation-only" argument as regulators have long believed of their authority to act and have acted to address spurious correlations.

While we urge the removal of "rational explanation" in the white paper, we also urge the CASTF to forcefully reject the false claim of legislative authority usurpation and to respond to the flawed NCOIL/Industry analysis regarding "correlation only."

2. The white paper's failure to address proxy discrimination against protected classes is a glaring omission and renders the white paper largely irrelevant.

In CEJ's November 22, 2019 comments, we suggested that the white paper better address proxy discriminating against protected classes.

- We suggested a clarification that unfair discrimination means both the absence of a cost-based relationship and proxy discrimination against protected classes in Section.

- We also suggested the addition to the regulatory review of data used in the development and validation of the model the following:

  Determine if data used for model development and testing are biased against protected classes of consumers, if insurers have tested the data for such bias and if any action has been taken to eliminate or reduce bias in data.

- We also suggested the addition of a section "Testing for and Minimizing Disparate Impact Unfair Discrimination."

All of CEJ's suggestions were intended to identify and minimize the impact of systemic racism and inherent bias in insurance within the cost-based foundation of insurance.

CASTF did not respond to the first or third comment. In response to the second comment, CASTF largely rejected our comments. In response to CEJ's comment on regulatory review of data for bias against protected classes, CASTF wrote:

> *Industry does not collect information that could clearly demonstrate if there is disparate impact on protected classes. It is beyond the scope of this paper to propose that the collection of data on protected classes is necessary in order to review models underlying rating plans.*

Further, the white paper included the assessment of proxy discrimination as an "other considerations" that "were not within the scope of this paper."

2.1 Proxy discrimination against protected classes must be a central focus of regulatory review and, consequently, of the white paper.

The use of predictive models in pricing is overwhelmingly oriented towards risk segmentation. While there are a few examples of predictive models used for estimating aggregate losses – catastrophe models – insurers use predictive models to develop and employ risk classification for underwriting, tier placement and rating. Traditional actuarial techniques are largely employed for overall rate need – the "not excessive" and "not inadequate" components of the rate standards. But predictive models are used – with few exceptions – for risk classification and are therefore subject to the "not unfairly discriminatory" rate standard.

With the understanding that the vast majority of predictive models used by insurers and reviewed by regulators for risk classification, the primary goal of regulatory review of these models is to identify and prevent unfair discrimination.

Given that the purpose of regulatory review of complex, predictive models is primarily – and overwhelmingly – to identify unfair discrimination, it is inconceivable that the white paper suggest that examining these models for unfair discrimination against protected classes is outside of the scope of the paper. CASTF has arbitrarily and mistakenly eliminated a core and crucial component of regulators' statutory responsibility by claiming examination of proxy discrimination against protected classes is outside the scope of the white paper.

2.2 Proxy discrimination against protected classes is a clear concept.

"Proxy discrimination against protected classes" is a clear concept for regulators. It means discriminating against a class of consumers identified for protection in statutes or regulation by means of a proxy for the class identification. Both parts of the phrase are well understood – protected classes and proxy discrimination.

Further, we know before the murder of George Floyd that proxy discrimination against protected classes is one of the major consumer protection issues with big data analytics and predictive models. We also know that addressing such algorithm bias has been at the core of every set of AI principles produced around the world, including the current draft of the NAIC Principles on AI.

However, the murder of George Floyd and the Black Lives Matter movement has put proxy discrimination resulting from systemic racism and inherent bias into stark relief. The issue of proxy discrimination against protected classes is at last as important as the issue of cost-based justification in regulatory review of predictive models.

2.3 Proactive efforts to identify and minimize proxy discrimination against protected classes is not just consistent with the cost-based foundation of insurance, but improves cost-based pricing.

There is a lengthy history of applying disparate impact unfair discrimination analysis in a variety of industries, including insurance. The federal Fair Housing Act has recognized proxy discrimination or disparate impact against protected classes for over 40 years and the history of disparate impact challenges shows that such challenges, when successful, improve cost-based pricing.

For example, in the 1990's, fair housing groups brought a disparate impact challenge against insurers' use of age and value of the home for underwriting. The groups argued that these underwriting guidelines discriminated against minority communities because these communities' housing was characterized by low value and old age. The challenges were largely successful and, in response, insurers developed more detailed underwriting based on, for example, age and type of electrical system and age and condition of the roof. The result was not just fairer treatment of minority communities, but improvements in cost-based pricing by insurers.

2.4 The fact that insurers do not collect protected class characteristics from applicants and policyholders is not an impediment to proactive efforts to identify and minimize proxy discrimination against protected classes.

As noted above, CASTF failed to address the core issue of proxy discrimination against protected classes in the white paper by, among other things, falsely equating insurers' non-collection of protected class characteristics with insurers' inability to analyze their data and models for proxy discrimination.

In fact, data and other tools are available for insurers to assign proxies for race to data records in order to test for proxy discrimination. In some instances, the tests will be simple. For example, in one of the CASTF book clubs, a vendor developing auto telematics pricing models stated they don't do anything to test for bias against protected classes. Yet, it would be easy for this vendor to examine whether the telematics data used in the development of the models reflects fair and unbiased availability of the telematics program across all communities – by simply mapping the garaging address to Census data – namely, the majority racial composition of census blocks. Yet, the vendor dismissed even this simple test of bias in data despite the well-known evidence that biased data reflect and perpetuate historic discrimination.

Further, valid statistical proxies for race / ethnicity are available and have been used to test for proxy discrimination against protected classes. For example, attached to our comment letter, please find:

Using publicly available information to proxy for unidentified race and ethnicity, Consumer Financial Protection Bureau, 2014

"Assessing Fair Lending Risks Using Race/Ethnicity Proxies." Yan Zhang, Office of the Comptroller of the Currency

"A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity," Elliot, et al, Heath Services Research, October 2008.

"Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination," Bogen, Rieke and Ahmed, 2020

These resources just touch the surface of possibilities for insurers to test for and minimize proxy discrimination against protected classes. If regulators were to routinely ask for – or include in the white paper as a routine part of the review of predictive models – insurer actions to detect and minimize proxy discrimination against protected classes, insurers – as well as data vendors and consulting firms – would develop the tools to fulfill this core analysis of unfair discrimination.

2.5 CEJ asks the CASTF to include of guidance regarding testing for and minimizing proxy discrimination against protected classes.

We ask CASTF to revise the white paper to include CEJ's suggested guidance regarding identification and minimization of proxy discrimination against protected classes. CASTF's rationales for exclusion – that protected class data are not available and that proxy discrimination against protected classes is outside the scope of the white paper – are demonstrably incorrect.

Data to perform analyses of proxy discrimination analysis are available, as noted above. The fact that the data may not be perfect is not a valid excuse. Data do not have to be perfect to be sufficiently valid to produce a reliable analysis.

Further, the issue of proxy discrimination against protected classes is directly related to CASTF's concern about "rational explanations." Just as proxy discrimination against protected classes represents a spurious correlation – as in the example above where the age and value of the home had a spurious correlation to claims – so does CASTF's concern with "rational explanation." The analytic tools to identify and minimize proxy discrimination against protected classes is a scientific, statistically-valid and objective approach to addressing the spurious correlation concern reflected in the "rational explanation" guidance. Pages 5 to 8 of the attached "CEJ's Call on Insurers and Regulates to Address Systemic Bias and Inherent Racism in Insurance" explains the scientific foundation of a disparate impact analysis.

Finally, given the NAIC's recent commitment to address race and diversity and the variety of actions that reflect this commitment, it would be contradictory for the CASTF to ignore the issues of systemic racism and inherent bias in a white paper providing guidance for regulatory review of predictive models.

Comments of the Center for Economic Justice

To NCOIL Regarding the Proposed

"Resolution Urging the National Association of Insurance Commissioners to Refrain from Intruding on the Constitutional Role of State Legislators."

June 28, 2020

The Center for Economic Justice (CEJ) suggests that NCOIL withdraw the ill-conceived "Resolution Urging the National Association of Insurance Commissioners to Refrain from Intruding on the Constitutional Role of State Legislators." The Resolution suffers from a number of false statements, fails to recognize the reality of current ratemaking and regulatory review, miscomprehends the oft-repeated term "correlation," represents an endorsement of proxy discrimination against protected classes and misdiagnoses the problem with the white paper's use of rational explanation. Among the problems with the resolution:

1. It is unclear why NCOIL has decided that a technical paper regarding review of complex pricing algorithms is the target for the proclamation of correlation as the intent and sole purview of state legislators. The fact that, among the many critical issues facing insurance consumers, NCOIL has prioritized an industry complaint feeds the perception by some that NCOIL's actions reflect the priorities of its industry corporate sponsors.

2. The premise of the resolution – "established rate filing review is based on correlation" – is demonstrably false and unsupported by statutory language. Neither of the NCOIL rating models cited in the resolution used the term "correlation." The purported reliance on "correlation-only" conflicts with the language of the NCOIL models regarding unfair discrimination.

3. As a former regulator charged with review and approval of rate filings and an expert witness in administrative and judicial proceedings on unfair discrimination and risk classification in insurance for nearly 30 years, simple correlation has never been sufficient justification for a risk classification.

4. The repeated references to "correlation" divorce the resolution from the reality of rate filings today. Insurers now use algorithms – whether for pricing, claims, anti-fraud or more – based on statistical techniques light years from simple correlation.

5. The repeated references to "correlation" are an endorsement of proxy discrimination. By declaring that any correlation is sufficient justification – even if that correlation is a proxy for discrimination against a protected class – and defending such proxy discrimination on the basis of states' rights the resolution ignores and repudiates the commitment and efforts by industry and regulators to address systemic racism in insurance.

6. The problem with the use of "rational explanation" in the CASTF White Paper is not a usurpation of legislative prerogative. Rather, "rational explanation" is a subjective approach to the problem of identifying spurious correlations.

Why This Resolution Targeting a NAIC Technical White Paper Now?

Insurance regulators at the NAIC have been grappling for over five years with the revolution in insurance operations resulting from insurers' use of big data analytics, complex algorithms, artificial intelligence and machine learning. The regulators' concerns are being examined in the NAIC's Artificial Intelligence Working Group, the Accelerated Underwriting Working Group, the Big Data Working Group, the Innovation and Technology Task Force, the Casualty Actuarial Task Force and more. Insurers' use of big data analytics represents a revolution in insurer operations that has challenged both regulators' ability to keep up with industry practices and for decades-old statutory authorities to provide the necessary consumer protections.

Of all the NAIC activities related to regulatory responses to insurers' use of big data analytics, it is curious that NCOIL has prioritized – in the current period of pandemic and systemic racism issues – with a phrase in a 50-page NAIC white paper – to proclaim a resolution. The fact that NCOIL chooses to prioritize this particular industry complaint about a NAIC technical white paper will fuel the contention of some that NCOIL parrots the interests of its industry corporate sponsors.

As discussed further below, the problem with the term "rational explanation" in the white paper, is not that it challenges state legislative authority, but that it is a technically incorrect approach to addressing problems of spurious correlations.

False Foundation – "Correlation" Does Not Appear in NCOIL and NAIC Rating Models

The foundation of resolution is that claim, "**WHEREAS**, established rate filing review is based on correlation, which demonstrates that rating variables are valid so long as they correlate with a loss."

Yet, the term "correlation" does not appear in either of the NCOIL rating models cited in the resolution. Nor does "correlation" appear in any of the NAIC property casualty rating

models.[1] Nor does "correlation" appear in the Casualty Actuarial Society's "Statement of Principles Regarding Property and Casualty Insurance Ratemaking."[2] Nor does it appear in the American Academy of Actuaries "Risk Classification Statement of Principles."[3] The term "correlative classes" appears once in the Risk Classification of Principles in a section on Credibility and not in the manner suggested by the resolution.[4] These risk classification principles identify a variety of considerations in developing risk classifications, including stability in avoiding abrupt changes in prices, maximizing the availability of coverage, minimizing ability to manipulate or misrepresent a risk characteristic and the need for public acceptability.

> Any risk classification system must recognize the values of the society in which it is to operate. This is a particularly difficult principle to apply in practice, because social values:
>
> - are difficult to ascertain;
> - vary among segments of the society; and
> - change over time.
>
> The following are some major public acceptability considerations affecting risk classification systems:
>
> - They should not differentiate unfairly among risks.
> - They should be based upon clearly relevant data.
> - They should respect personal privacy.
> - They should be structured so that the risks tend to identify naturally with their classification.

In fact, a simple "correlation" is not the basis for fair discrimination. NAIC models define unfair discrimination to exist if "after allowing for practical limitations, price differentials fail to reflect equitably the differences in expected losses and expenses." The NCOIL models don't define unfair discrimination other than discrimination "on the basis of race, color, creed, or national origin."

If, as claimed in the resolution that "rate filing review is based on correlation," then the appropriate test for discriminating "on the basis of race, color, creed, or national origin" would also be a simple correlation between the rating factor and the prohibited classifications.

---

[1] See https://www.naic.org/store/free/GDL-1780.pdf and https://www.naic.org/store/free/GDL-1775.pdf and https://www.naic.org/store/free/GDL-1781.pdf

[2] https://www.casact.org/professionalism/standards/princip/sppcrate.pdf

[3] http://www.actuarialstandardsboard.org/wp-content/uploads/2014/07/riskclassificationSOP.pdf

[4] "Accurate predictions for relatively small, narrowly defined classes often can be made by appropriate statistical analysis of the experience for broader groupings of correlative classes.

Miscomprehension of "Correlation" and Regulatory Review of Rate Filings

The resolution incorrectly equates simple correlation with the statutory standards for rates. A correlation is simply the extent to which a pair of variables are related. There are many correlations between variables that bear no relationship to one variable predicting the other variable – and that latter is the essence of a rating factor identifying price differentials among consumers in the cost of the transfer of risk.

Here are some examples of very highly correlated variables, which are also examples of "spurious correlation"[5] – "two or more events or variables that are associated but not causally related due to either coincidence or a third unseen factor."[6] A perfect correlation is 100%. No correlation is 0%.

- There was a 94.7% correlation between per capita cheese consumption and the number of people who dies by becoming tangled in their bedsheets from 2000 to 2009.
- There was a 99.3% correlation between the divorce rate in Maine and per capita consumption of margarine from 2000 to 2009. As an aside, the Indiana Department of Insurance disapproved a rate filing in which the insurer sought to use per-capita margin consumption as a risk classification.
- There was a 98.5% correlation between total revenue generated by arcades and computer science doctorates awarded in the US from 2000 to 2009.

In the 30 years that I have been reviewing rate filings and risk classifications and regulatory activity in this arena, a simple correlation has never been a sufficient justification for a rating factor.

We offer two real life examples to demonstrate why this is the case. First, in the early 1990s in Texas, an insurer in Texas sought approval for a homeowners discount based on tenure with insurer – if an insured was with the company for several years, they would bet a discount. The insurer provided the following information[7]:

Tenure (Years)	1	2	3	4	5	6	7	8
Loss Ratio	64.0%	63.4%	62.8%	62.2%	61.6%	61.0%	60.4%	60.0%

[5] https://tylervigen.com/spurious-correlations
[6] https://en.wikipedia.org/wiki/Spurious_relationship
[7] These are not the actual numbers, but an illustration of the actual situation.

Based on this simple "correlation," the loss ratio seemed to track years of tenure with the company. By the standards of the resolution, this presentation of loss ratios would have been the end of discussion and prohibited any further inquiry by the regulator. In fact, the company was asked to produce loss ratios by years of tenure separate for homeowners (e.g., HO-3) policies and renters' policies (e.g. HO-4). It turned out that the company had combined the experience.

When looked at separately, the loss ratios for each of the two types of policies didn't vary with tenure. Homeowners loss ratios were consistent and consistently lower than those for renters' policies. The spurious findings in the table above were a result of the percentage of renters' policies declining as a share of total homeowners policies over time – far fewer people rent for five, six, or seven years than for one or two years so the declining loss ratios in the table were a result of fewer high-loss ratio renters' policies for each additional year of tenure.

A second example comes from a disparate impact challenge under the federal Fair Housing Act. In the mid 1990s, fair housing groups challenged insurers' use of age and value of the home as underwriting factors for homeowners insurance. The insurers used these factors because of a correlation to expected losses. The fair housing groups showed that using age and value of the home served as proxies for race and income. Because of historical discrimination in housing and mortgages, the housing in communities of color was characterized by older age and lower values. When confronted with the data, the insurers recognized they were using a proxy for condition of the home that was, in fact, a proxy for race. The insurers stopped using age and value of the home and started using more accurate variables like age and condition of the roof and type of electrical system. By responding to the disparate impact challenge, insurers stopped penalizing minority homeowners who maintained their homes with race-based underwriting.

Miscomprehension of Insurer Rating Practices and the Challenges for Regulators

The resolution's references to "correlation" seem like a quaint reference to a long-gone – by 50 years – era. The same NAIC Casualty Actuarial Task Force holds monthly "book clubs" in which insurers and experts make presentations on current ratemaking practices. This past week was an example in which Allstate subsidiary Arity made a presentation on the development of their telematics pricing models for auto insurance.[8] The title of the presentation was "Modeling concepts, hyperparameter tuning, and telematics." The presentation reviewed the parts of a scoring (pricing) model, including ordinary least squares regression, generalized linear models, generalized linear models with log link functions, decision tree models, neural nets, gradient descent, hyperparameters and extreme gradient boosting. Needless to say, that when a regulator is presented with rating factors based on such a model, it is meaningless to try to look for a simple correlation.

[8] https://content.naic.org/sites/default/files/call_materials/Modeling%20concepts%20hyperparameter%20tuning%20and%20telematics.pdf

It is this new and massive complexity – actuarial science merged with data science merged with astrophysics – that presents the challenge for regulators to enforce current statutes. We suggest that instead of a resolution harkening back to a by-gone era that never really existed, NCOIL's efforts would be better spent working with regulators to modernize regulatory authorities and capabilities to deal with the reality of complex models in insurance.

A challenge for insurers and regulators that has always existed and continues to exist is whether a particular relationship – correlation – is real or spurious. When insurers have tried to utilize the closest thing to a simple correlation, insurers and regulators have found problems. Thirty or more years ago, insurers may have presented justification for a particular rating factor with what is known as a univariate analysis – comparing one predictive variable to, say, loss ratio. With traditional actuarial practices, looking at two or more variables at the same time was difficult because each additional variable required more data for a credible – or reliable – analysis. But the univariate analysis always had problems because insurers and regulators knew that, in addition to any correlation between particular rating factors and loss ratio, there was correlation between the rating factors with the result that univariate analysis led to double counting.

For example, both age and miles driven are related to expected losses. But as drivers get older – and retire form work – they drive less. So, a simple analysis of age and expected losses is reflecting the correlation miles driven and vice versa. So using both based on independent analyses yields double counting.

Since the early 1990s – at least – insurers have moved to new statistical techniques to develop and analyze rating factors. These techniques permit the simultaneous analysis of multiple variables and remove the correlation among the variables to eliminate double counting of impact on outcomes. Stated differently, the multivariate techniques used today advance from and address the limitations of "correlation."

This issue is discussed in greater detail in the attached "CEJ Call to Insurers and Insurance Regulators to Address Systemic Racism in Insurance."

Tacit Endorsement of Proxy Discrimination against Minority Consumers and Other Protected Classes in the name of States' Rights.

The repeated references to "correlation" in the resolution are an endorsement of proxy discrimination. By declaring that any correlation is sufficient justification – even if that correlation is a proxy for discrimination against a protected class and defending such proxy discrimination on the basis of states' rights – ignores the commitment and efforts by industry and regulators to address systemic racism in insurance.

By the standard espoused in the resolution, a rating factor that was a proxy for being a Black American is legitimate as long as there is a correlation to losses. Never mind that the factor is a proxy for a prohibited class or that that the factor discriminates on the basis of a prohibited factor.

Some data vendors offer a criminal history score that purports to score homeowners insurance on the basis of complaints filed with courts. Based on the resolution, as long as there was a "correlation," that would not only be okay, but regulators are prohibited from further inquiry. What would the use of a criminal history score look like in the case of George Floyd, if he lived? What would the use of a criminal history score look like in Ferguson, Missouri, where the US Department of Justice found the following.

> US DOJ Investigation of the Ferguson Police Department
> Ferguson's approach to law enforcement both reflects and reinforces racial bias, including stereotyping. The harms of Ferguson's police and court practices are borne disproportionately by African Americans, and there is evidence that this is due in part to intentional discrimination on the basis of race.
>
> Ferguson's law enforcement practices overwhelmingly impact African Americans. Data collected by the Ferguson Police Department from 2012 to 2014 shows that African Americans account for 85% of vehicle stops, 90% of citations, and 93% of arrests made by FPD officers, despite comprising only 67% of Ferguson's population.
>
> FPD appears to bring certain offenses almost exclusively against African Americans. For example, from 2011 to 2013, African Americans accounted for 95% of Manner of Walking in Roadway charges, and 94% of all Failure to Comply charges.
>
> Our investigation indicates that this disproportionate burden on African Americans cannot be explained by any difference in the rate at which people of different races violate the law. Rather, our investigation has revealed that these disparities occur, at least in part, because of unlawful bias against and stereotypes about African Americans.

It would be interesting to count the number of NCOIL members who have received citations for Manner of Walking in Roadway, let alone been penalized with higher insurance rates as a result.

In the aftermath of the murder of George Floyd, many insurer CEOs made statements declaring their personal and corporate opposition to inherent bias and systemic racism. The NCOIL resolution goes in the other direction – it defends systemic racism in insurance by prohibiting inquiry into proxy discrimination. This unfortunate position by NCOIL is also tone-deaf. It relies upon the same states' rights argument used by those opposing the abolition of slavery and integration.

The Problem with the White Paper's Use of "Rational Explanation" is Not a Challenge to Statutory Standards, but a Technical Issue with Identifying Spurious Correlations

The CASTF's white paper use of "rational explanation" is problematic because it is a subjective approach to addressing spurious correlations. It is not a challenge to the mythical statutory standards in the resolution because regulators and actuarial standards of practice have always sought to distinguish between real and false relationships among predictive variables in insurance. "Rational explanation" is problematic because "rational" is subjective – a rational explanation to one person may not be rational to another. The way to address the problem with "rational explanation" is to urge regulators to utilize more of the advanced analytic and statistical tools to distinguish between fair and proxy discrimination. Again, the attached CEJ paper discusses this in more detail.

> The NAIC Casualty Actuarial and Statistical Task Force deals generally with actuarial issues in property casualty lines of insurance. The Task Force is currently developing a white paper to provide best practices for regulatory review of complex pricing models used by insurers to justify rates. The current draft does not incorporate identification and minimization of systemic bias or disparate impact, but simply lists it as another consideration. Insurance rate standards include rates not being excessive, not being inadequate and not being unfairly discriminatory.
>
> The use of complex predictive models for pricing by insurers is focused on risk segmentation and the development of risk classifications and rating factors. Traditional actuarial techniques – not complex predictive models – are generally used for overall rate level indications – the metric for assessing whether rates are excessive or inadequate. The overwhelming reason for close scrutiny of complex predictive models by regulators is to assess whether the risk classifications are fair or unfairly discriminatory. It is an understatement to say that the current draft white paper has a massive whole because of the failure to address proxy discrimination and disparate impact. Guidance to insurance regulators for regulatory review of complex insurance predictive models should prioritize the identifications and minimization of systemic bias and disparate impact.

Conclusion

For a myriad of reasons, CEJ suggests that NCOIL withdraw this deeply-flawed resolution.

CENTER for ECONOMIC JUSTICE | FAIR ACCESS FAIR TREATMENT

The Center for Economic Justice's Call to Insurers and Insurance Regulators

To Address Societal Systemic Bias and Inherent Racism in Insurance By Explicit Recognition of Disparate Impact as Unfair Discrimination in Insurance

Submitted to the National Association of Insurance Commissioners'
Big Data Working Group
Artificial Intelligence Working Group
Market Regulation and Consumer Affairs Committee
Casualty Actuarial and Statistical Task Force
Accelerated Underwriting Working Group

June 18, 2020

Action, Not Just Words, Needed

The murder of George Floyd has led to widespread corporate recognition of and opposition to systemic bias and inherent racism in America. Corporate CEOs have spoken out, including major insurer CEOs.

> "In the coming days, I encourage each of us to step outside of our comfort zones, seek to understand, engage in productive conversations and hold ourselves accountable for being part of the solution. We must forever stamp out racism and discrimination." Those are the words of Kirt Walker, Chief Executive Officer of Nationwide.
>
> Floyd's death in Minneapolis is the latest example of "a broken society, fueled by a variety of factors but all connected by inherent bias and systemic racism. Society must take action on multiple levels and in new ways. It also requires people of privilege—white people—to stand up for and stand with our communities like we never have before," Those are the words of Jack Salzwedel, the CEO of American Family.

Perhaps this will be a turning point in insurer and regulatory practices, **but insurers have consistently opposed proposals to address systemic bias and inherent racism in insurance.** This opposition has come in two general themes – opposition to any responsibility by insurers or regulators to identify and minimize **disparate impact**[1] in insurance and opposition to any form of regulatory data collection to allow regulators and the public to assess market outcomes and thereby hold insurers accountable for their practices.

While insurers have been constant in opposing any responsibility to address systemic bias and inherent racism – in contrast to the recent public statements of insurer CEOs – most state insurance regulators believe they have the authority to stop proxy discrimination against protected classes. This belief, however, has never manifested itself, in regulatory standards, models laws or consistent approaches across states.

If insurers and insurance regulators truly want to address systemic bias and inherent racism in insurance, two long-overdue actions are needed.

1. Explicit recognition of disparate impact as unfair discrimination against protected classes in insurance coupled with responsibility for insurers and insurance regulators to identify such disparate impact and take steps to minimize this proxy discrimination within the overall regulatory framework of cost-based pricing.

2. Development of regulatory data collection and analysis infrastructure and capabilities for insurance regulators and the public to meaningfully monitor market outcomes for all consumers, to identify discriminatory outcomes and trace disparate impact to the causes.

[1] Disparate impact refers to practices that have the same effect as disparate treatment or intentional discrimination against protected classes. Protected classes reefer to those consumer characteristics which may not be the basis for discrimination and include, in most states, race, religion and national origin. Disparate impact is also known as disparate effect or proxy discrimination – discrimination against the protected class through a proxy for the protected class characteristic. Disparate impact as unfair discrimination has long been recognized under federal employment and housing laws. In 2015, the U.S. Supreme Court affirmed disparate impact as unfair discrimination under the Fair Housing Act which covers a variety of housing-related issues, including insurance, with Justice Kennedy writing, "Recognition of disparate-impact liability under the FHA plays an important role in uncovering discriminatory intent: it permits plaintiffs to counteract unconscious prejudices and disguised animus that escape easy classification as disparate treatment."

The mechanisms to accomplish these actions are straightforward.

1. Development of, and implementation by the states, through the National Association of Insurance Commissioners (NAIC)[2] of a model law addressing algorithmic bias including recognition of disparate impact as unfair discrimination against protected classes in insurance with guidance and safe harbors for insurers to identify and minimize disparate impact in marketing, pricing, claims settlement and anti-fraud efforts.

2. Development of, and implementation by the states, through the NAIC, of a market regulation data collection and analysis infrastructure to timely and meaningfully monitor consumer insurance outcomes – similar in scope and capability to what state insurance regulators and the NAIC currently have for monitoring the financial condition of insurers.

In the absence of the necessary actions by insurers and the states, Congress and federal agencies will eventually address these problems through civil rights legislation and enforcement.

In An Era of Big Data Analytics and Insurers' Rapidly Growing Use of Third-Party Data and Complex Algorithms, the Potential For Algorithmic Bias and Proxy Discrimination Has Grown Dramatically.

The potential for big data, artificial intelligence, machine learning – implemented through rapid deployment of complex algorithms – has increased the potential for intentional or unintentional proxy discrimination through algorithmic bias. This potential is well recognized. Barocas and Selbst state the issue succinctly:[3]

> Advocates of algorithmic techniques like data mining argue that they eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data mining can inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large. Often, the "patterns" it discovers are simply preexisting societal patterns of inequality and exclusion. Unthinking reliance on data mining can deny members of vulnerable groups full participation in society.

Most data sets of personal consumer information as well data sets of the built environment reflect historical discrimination against protected classes. For example, TransUnion has an insurance score used for pricing based on criminal violations filed with the courts – not just convictions, but all criminal filings regardless of the eventual outcome. TransUnion's marketing materials state:

[2] https://content.naic.org/index_about.htm
[3] http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

"TransUnion recently evaluated the predictive power of court record violation data (including criminal and traffic violations)

"Also, as court records are created when the initial citation is issued, they provide insight into violations beyond those that ultimately end up on the MVR—such as violation dismissals, violation downgrades, and pre-adjudicated or open tickets."

It did not take the recent murders of Black Americans by police to recognize that this "criminal history score" will reflect historic discrimination in policing against Black Americans and perpetuate that discrimination in insurance. Consider policing records in Ferguson, Missouri.

US DOJ Investigation of the Ferguson Police Department
Ferguson's approach to law enforcement both reflects and reinforces racial bias, including stereotyping. The harms of Ferguson's police and court practices are borne disproportionately by African Americans, and there is evidence that this is due in part to intentional discrimination on the basis of race.

Ferguson's law enforcement practices overwhelmingly impact African Americans. Data collected by the Ferguson Police Department from 2012 to 2014 shows that African Americans account for 85% of vehicle stops, 90% of citations, and 93% of arrests made by FPD officers, despite comprising only 67% of Ferguson's population.

FPD appears to bring certain offenses almost exclusively against African Americans. For example, from 2011 to 2013, African Americans accounted for 95% of Manner of Walking in Roadway charges, and 94% of all Failure to Comply charges.

Our investigation indicates that this disproportionate burden on African Americans cannot be explained by any difference in the rate at which people of different races violate the law. Rather, our investigation has revealed that these disparities occur, at least in part, because of unlawful bias against and stereotypes about African Americans.

One of the oft-cited benefits of big data analytics in insurance is greater personalization – the ability of insurers to develop products and pricing tailored to individual needs and characteristics. But the other side of personalization is exclusion. Insurers' use of algorithmic techniques called price optimization, claim optimization and customer lifetime value are examples of the flip side of big data personalization – differential treatment of groups of consumers that reflect and perpetuate inherent bias and systemic racism.

The TransUnion Criminal History Score is just one example – egregious and obvious – of algorithms that reflect and perpetuate historic discrimination against protected classes in insurance – algorithms that reinforce inherent bias and systemic discrimination. Others include:

- Employment categories and education levels for marketing, underwriting and pricing
- Price Optimization and Customer Lifetime Value Algorithms used for marketing, underwriting, pricing and claims settlement
- Facial analytics used in life insurance underwriting
- Household composition used for underwriting and pricing
- Credit scores for marketing, underwriting, pricing, claims settlement and anti-fraud efforts
- Fraud detection models based on biased learning data

Many of these practices have shown to discriminate unfairly against protected classes, generally, and Black Americans, specifically. A number of cities – as well as Google and IBM – have stopped using facial recognition technology because of the biases against Black Americans. After the New York Department of Financial Services developed a regulation permitting the use of employment and education characteristics in auto insurance pricing only if the insurer could demonstrate the practice did not unfairly discriminate against protected classes, insurers' use of the "risk" characteristics disappeared.

The Consumer Federation of America has produced a number of extraordinary studies of discriminatory market outcomes resulting from rating factors that reflect systemic racism.[4] Insurance industry trade associations have dismissed the CFA's discriminatory findings with the claim that insurers engage in cost-based, race-neutral practices – while refusing to both provide the data to back up these claims and refusing to recognize that systemic racism will show up as disparate impact.

If insurers and insurance regulators are serious about addressing inherent bias and systemic racism in insurance, then action is needed. Fortunately, the insurance industry has the precise skill set needed to identify and minimize disparate impact and insurance regulators have the resources to develop the necessary guidance and infrastructure.

Disparate Impact Analysis is Straightforward and Particularly Suited to Insurance.

The mechanics of a disparate impact analysis in insurance are straightforward and use well-accepted statistical and actuarial methods. Any algorithm – whether for pricing, anti-fraud, claims settlement, lifetime customer value, price optimization or other – takes the basic form of an equation in which certain variables or factors – the explanatory factors – seek to explain or predict a particular outcome.

[4] https://consumerfed.org/cfa-studies-on-the-plight-of-low-and-moderate-income-good-drivers-in-affording-state-required-auto-insurance./

Consider the following general model.

$$b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e = y$$

Say that $X_1, X_2 + X_3$ are explanatory variables used to predict y – the frequency of an auto claim, for example.

Let's assume that all three Xs are statistically significant predictors of the likelihood of a claim and the b values associate with each X are how much each X contributes to the explanation of claim.

b_0 is the "intercept" – a base amount and e is the error term – the portion of the explanation of the claim not provided by the independent variables.

When the algorithm or model is developed, the modeler will typically data mine some database of personal consumer information, built environment or natural environment for characteristics that are correlated with the desired outcome. These variables are combined into a model, but a variable that might be predictive on its own can lose its predictive capability when combined with other variables because the variables might be correlated with one another. In that event, the variable serving as the proxy for the other variable loses its individual explanatory power. In our example, above, if, say, $X_1 + X_2$ are highly correlated, when the two variables are used in the same algorithm, one of the variables will lose its predictive power.

From a statistical and actuarial perspective, a disparate impact analysis does two things. First, it examines the amount of correlation between explanatory variables or factors and protected class characteristics to determine if any of the explanatory variables have significant correlation with, and thereby serve as proxies – in whole or in part – for protected class characteristics.

The second function of a disparate impact analysis is to remove the correlation between the explanatory variables and protected class characteristics with the result that the remaining explanatory power of the explanatory variables is the independent contribution – independent of correlation to protected class characteristics – of the explanatory variables relationship to the outcome.

Consider the following example. Suppose an explanatory factor was perfectly correlated with being a Black American. In statistical terms, this means a perfect or 100% correlation and the explanatory factor is a perfect proxy for being African-American. Assume that when used in an algorithm, this perfect proxy for being a Black American shows us as predictive of some outcome variable. Assume variable X_1 in our simple model above is the perfect proxy characteristic and further assume that the proxy variable shows a correlation to / is predictive of the outcome variable. Given our assumption that variable X_1 is a perfect proxy for being Black American, then the results of the model would be identical whether we used the proxy variable or used Black American explicitly. If the proxy variable is used, this would not be intentional discrimination – defined as explicit use of a protected class characteristic – even though it has

precisely the same effect. While most regulators believe they have the authority and obligation to stop the use of such proxies for protected class characteristics, the insurance industry view, as espoused by the American Property Casualty Insurance Association, is that even in this extreme case, there is no unfair discrimination against a protected class.

When the data are run through the model, variable X_1 shows some correlation to the outcome variable and is, therefore, "predictive." But, what it is really doing is simply standing in for being Black American and indirectly discriminating on the basis of race. This proxy factor is, in fact, simply reflecting and perpetuating discrimination against Black Americans.

One approach to disparate impact analysis – among many which generally try to remove the correlation between predictive variables and protected class characteristics – is to include a control explanatory variable for being Black American in the algorithm. Let's know add a new variable to algorithm – a specific variable for being Black American.

$$b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \mathbf{b_4R_1} + e = y$$

In statistical and actuarial terms, this is known as adding a control variable. The purpose of the control variable is to remove known correlations and biases in the other explanatory variables in order to better assess the independent and unique explanatory power of these other explanatory variables. For example, in personal auto pricing models, an insurer developing a national pricing model will utilize a control variable for State to remove the effects of / correlations with other explanatory variables of State-specific characteristics, such as different minimum liability limits, different tort and no-fault systems and different population distributions by age or other factors, among other things. In our example, our control variable $\mathbf{R_1}$ is being Black American.

Now, when the data are run through the model, explanatory variable X_1 – the perfect proxy for being Black American – shows no explanatory power and the control variable now shows the explanatory power that explanatory variable X_1 had in the original model. This is statistical evidence that explanatory variable X_1 was discriminating on the basis of race.

Let's consider two other examples – one in which there is a 50% correlation between variable X_1 and being Black American and a second in which there is a 0% correlation. In the 50% correlation, the variable X_1 may still show up as predictive of the outcome, but that predictive power will be different than from our first model without the control variable for being Black American. X_1's new contribution to explaining or predicting the outcome will now be its contribution independent of any correlation to being Black American. Consequently, disparate impact is recognized and minimized. Again, this is a common statistical and actuarial technique.[5]

[5] For example, the technique is explained is the chapter, ""Credit Scoring and the Fair Lending Issue of Disparate Impact, "in *Credit Scoring for Risk Managers*, Elizabeth May, editor, 2004.

In our third example there is 0% correlation between the variables X_1 and $\mathbf{R_1}$. In this situation, the predictive power of X_1 remains the same as in the original model because there is no disparate impact.

As noted above, disparate impact analysis is particularly suited to insurance because the actuarial justification required for insurance risk classifications is a statistical test – is the characteristic correlated with risk of loss? The same statistical test can be used to evaluate and minimize disparate impact. Stated differently – if a particular correlation and statistical significance is used to justify, say, insurance credit scoring, those same standards of correlation and statistical significance are reasonable evidence of disparate impact and unfair discrimination on the basis of prohibited factors.

In addition, the ability of insurers to identify and minimize disparate impact can be easily built into the development of pricing, marketing or claim settlement models by including consideration of prohibited characteristics as control variables in the development of the model and then omitting these prohibited characteristics when the model is deployed. Again, this is one of many ways to remove the correlations between explanatory variables in algorithms and the protected class characteristics that result in reflection of and perpetuation of historic discrimination or disparate impact.

Recognition by regulators and insurers of disparate impact as unfair discrimination in insurance against protected classes and requirements to identify and

- Minimizes Disparate Impact – Stop the Cycle of Perpetuating Historical Discrimination.
- Promotes Availability and Affordability for Underserved Groups
- Improves Cost-Based Insurance Pricing Models
- Improve Price Signals to Insureds for Loss Mitigation Investments
- Help Identify Biases in Data and Modelers / Improve Data Insights
- Improve Consumer Confidence of Fair Treatment by Insurers

What NAIC Committees and Working Groups Should Be Doing

The NAIC has spread work streams related to Big Data Analytics over a number of groups. With the exception of the Artificial Intelligence Working Group, none of these groups' work efforts address systemic bias in insurance.

Artificial Intelligence Working Group

The NAIC Artificial Intelligence (AI) Working Group is developing insurance-specific principles for the governance and use of AI in insurance. While there are a number of consumer protection issues associated with insurers' use of AI (or Big Data Analytics, generally), such as protection of personal data and transparency and accountability to consumers and regulators, the most important consumer protection is establishing a responsibility for insurers and regulators to identify and minimize algorithmic bias and proxy discrimination. Recognition of disparate impact and responsibility of insurers and regulators to minimize such systemic bias must be a core AI insurance principle.

Big Data Working Group

The NAIC Big Data Working Group is examining big data analytics issues across a variety of insurance operations and lines of business. The two actions called for by CEJ regarding disparate impact and data collection should be at the core of all the working group's inquiries and activities. The Big Data Working Group should be developing the model law or revisions to existing model laws regarding explicit recognition of disparate impact, guidelines for identify and minimizing proxy discrimination and safe harbors for insurers.

Market Regulation and Consumer Affairs Committee

The NAIC Market Regulation and Consumer Affairs Committee is the parent committee for a number of working groups related to insurance market regulation, including data collection for market regulation, market surveillance, market conduct examinations and antifraud efforts. The Committee should be a contributor to the development of model laws regarding disparate impact, but must take the lead on market regulation data collection – both to identify the types of data and algorithms used by insurers and what these data are used for and to re-engineer market regulation data collection to match the granularity and frequency of financial regulation data collection.

Casualty Actuarial and Statistical Task Force

The NAIC Casualty Actuarial and Statistical Task Force deals generally with actuarial issues in property casualty lines of insurance. The Task Force is currently developing a white paper to provide best practices for regulatory review of complex pricing models used by insurers to justify rates. The current draft does not incorporate identification and minimization of systemic bias or disparate impact, but simply lists it as another consideration. Insurance rate standards include rates not being excessive, not being inadequate and not being unfairly discriminatory.

The use of complex predictive models for pricing by insurers is focused on risk segmentation and the development of risk classifications and rating factors. Traditional actuarial techniques – not complex predictive models – are generally used for overall rate level indications – the metric for assessing whether rates are excessive or inadequate. The overwhelming reason

for close scrutiny of complex predictive models by regulators is to assess whether the risk classifications are fair or unfairly discriminatory. It is an understatement to say that the current draft white paper has a massive whole because of the failure to address proxy discrimination and disparate impact. Guidance to insurance regulators for regulatory review of complex insurance predictive models should prioritize the identifications and minimization of systemic bias and disparate impact.

Accelerated Underwriting Working Group

The NAIC Accelerated Underwriting Working Group continues the NAIC's multi-year examination of life insurers' use of Big Data analytics and predictive models in place of traditional actuarial practices for underwriting and pricing life insurance. While the predictive models now used by life insurers have the same function as those used in auto, home and other property casualty lines of insurance – namely, using non-traditional data and an algorithm to predict claims (or other outcomes of value to the insurer). While there are requirements for property casualty insurers to file these predictive models for regulatory review for some purposes – justifying rates – and special laws and provisions governing property casualty insurers' use of consumer credit information, there are no similar regulatory requirements for life insurers. The time is long overdue for this working group to develop the model laws for regulatory guidance and consumer protections to ensure consumer protections in the face of life insurers' growing use of non-traditional, non-insurance data and complex algorithms. And the core of such models laws and regulatory guidance must be identification and minimization of disparate impact and systemic racism.

Conclusion

Recent events have highlighted a long-standing gaps in insurer and insurance regulatory practices – the failure to monitor consumer market outcomes for discriminatory impacts against protected classes and the failure to incorporate identification and minimization of proxy discrimination in insurers' development of predictive models for all aspects of their operations and regulators' review of these algorithms. The tools are available to address these problems – analysis of disparate impact and improved data collection. CEJ calls on insurers and regulators to match their statements of outrage over systemic racism with the actions needed to identify and minimize such unfair discrimination in insurance.

Using publicly available information to proxy for unidentified race and ethnicity

A methodology and assessment

Summer 2014

Table of contents

1. Executive summary

The Consumer Financial Protection Bureau (CFPB) is charged with ensuring that lenders are complying with fair lending laws and addressing discrimination across the consumer credit industry. Information on consumer race and ethnicity is required to conduct fair lending analysis of non-mortgage credit products, but auto lenders and other non-mortgage lenders are generally not allowed to collect consumers' demographic information. As a result, substitute, or "proxy" information is utilized to fill in information about consumers' demographic characteristics. In conducting fair lending analysis of non-mortgage credit products in both supervisory and enforcement contexts, the Bureau's Office of Research (OR) and Division of Supervision, Enforcement, and Fair Lending (SEFL) rely on a Bayesian Improved Surname Geocoding (BISG) proxy method, which combines geography- and surname-based information into a single proxy probability for race and ethnicity. This paper explains the construction of the BISG proxy currently employed by OR and SEFL and provides an assessment of the performance of the BISG method using a sample of mortgage applicants for whom race and ethnicity are reported. Research has found that this approach produces proxies that correlate highly with self-reported race and national origin and is more accurate than relying only on demographic information associated with a borrower's last name or place of residence alone. The Bureau is committed to continuing our dialogue with other federal agencies, lenders, advocates, and researchers regarding the methodology.

2. Introduction

The Equal Credit Opportunity Act (ECOA) and Regulation B generally prohibit a creditor from inquiring "about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction"[1] with a few exceptions, including for applications for home mortgages covered under the Home Mortgage Disclosure Act (HMDA).[2] Information on applicant race and ethnicity, however, is often required to conduct fair lending analysis to identify potential discriminatory practices in underwriting and pricing outcomes.[3]

Various techniques exist for addressing this data problem. Demographic information that reflects applicants' characteristics—for example, whether or not an individual is White—can be approximated by constructing a proxy for the information. A proxy may definitively assign a characteristic to a particular applicant—an individual is classified as being either White or non-White—or may yield an assignment that is probabilistic—an individual is assigned a probability, ranging from 0% to 100%, of being White. When characteristics are not reported for an entire population of individuals, as is usually the case for non-mortgage credit products, techniques focused on approximating the demographic data generally require relying on additional sources of data and information to construct proxies.

---

[1] 12 C.F.R. § 1002.5(b).

[2] 12 C.F.R. § 1002.5(a)(2) and 12 C.F.R. § 1002.13. For HMDA and its implementing regulation, Regulation C, see 29 U.S.C § 2801-2810 and 12 C.F.R. Part 1003. For the Regulation B provisions concerning requests for information generally, see 12 C.F.R. § 1002.5.

[3] The ECOA makes it unlawful for "any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction (1) on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract); (2) because all or part of the applicant's income derives from any public assistance program; or (3) because the applicant has in good faith exercised any right under the Consumer Credit Protection Act." 15 U.S.C. § 1691(a).

3. Using census geography and surname data to construct proxies for race and ethnicity

In a variety of settings, including the analysis of administrative health care data and the evaluation of fair lending risk in non-mortgage loan portfolios, researchers, statisticians, and financial institutions often rely on publicly available demographic information associated with an individual's surname and place of residence from the U.S. Census Bureau to construct proxies for race and ethnicity when this information is not reported. A proxy for race and ethnicity may be based on the distribution of race and ethnicity within a particular geographic area. Similarly, a proxy for race and ethnicity may be based on the distribution of race and ethnicity across individuals who share the same last name. Traditionally, researchers and statisticians have relied on information associated with either geography or surnames to develop proxies.[4]

A research paper by Elliott et al. (2009) proposes a method to proxy for race and ethnicity that integrates publicly available demographic information associated with surname and the geographic areas in which individuals reside and generates a proxy that is more accurate than those based on surname or geography alone.[5] The method involves constructing a probability of

[4] For example, in conducting fair lending analysis of indirect auto lending portfolios, the Federal Reserve relies on the U.S. Census Bureau's Spanish Surname List to proxy for Hispanic borrowers. Information on the Federal Reserve's methodology is available at: http://www.philadelphiafed.org/bank-resources/publications/consumer-compliance-outlook/outlook-live/2013/indirect-auto-lending.cfm.

[5] Marc N. Elliott et al., Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities, HEALTH SERVICES & OUTCOMES RESEARCH METHODOLOGY (2009) 9:69-83.

assignment to race and ethnicity based on demographic information associated with surname and then updating this probability using the demographic characteristics of the census block group associated with place of residence. The updating is performed through the application of a Bayesian algorithm, which yields an integrated probability that can be used to proxy for an individual's race and ethnicity. Elliott et al. (2009) refer to this method as Bayesian Improved Surname Geocoding (BISG).

The Office of Research (OR) and the Division of Supervision, Enforcement, and Fair Lending (SEFL) employ a BISG proxy methodology for race and ethnicity in our fair lending analysis of non-mortgage credit products that relies on the same public data sources and general methods used in Elliott et al. (2009).[6] The following sections describe these public data sources, explain the construction of the BISG proxy, identify any differences from the general methods used by Elliott et al. (2009), and provide an assessment of the performance of the BISG proxy.

Statistical analysis based on proxies for race and ethnicity is only one factor taken into account by OR and SEFL in our fair lending review of non-mortgage credit products. This paper describes the methodology currently employed by OR and SEFL but does not set forth a requirement for the way proxies should be constructed or used by institutions supervised and regulated by the CFPB.[7] Finally, our proxy methodology is not static: it will evolve over time as enhancements are identified that improve accuracy and performance.

---

[6] We also rely on a proxy for sex based on publicly available data from the Social Security Administration, available at: http://www.ssa.gov/oact/babynames/limits.html. The focus of this paper, however, is on the BISG methodology and the construction of the proxies for race and ethnicity.

[7] The federal banking regulators have made it clear that proxy methods may be used in fair lending exams to estimate protected characteristics where direct evidence of the protected characteristic is unavailable. The CFPB adopted the Interagency Fair Lending Examination Procedures as part of its *CFPB Supervision and Examination Manual*. See CFPB Supervision and Examination Manual, Part II, C, ECOA, Interagency Fair Lending Examination Procedures at 19, available at http://files.consumerfinance.gov/f/201210_cfpb_supervision-and-examination-manual-v2.pdf (explaining that "[a] surrogate for a prohibited basis group characteristic may be used" in a comparative file review and providing examples of surname proxies for race/ethnicity and first name proxies for sex).

3.1 Data sources

3.1.1 Surname

Information used to calculate the probability of belonging to a specific race and ethnicity given an individual's surname is based on data derived from Census 2000 that was released by the U.S. Census Bureau in 2007.[8] This release provides each surname held by at least 100 enumerated individuals, along with a breakdown of the percentage of individuals with that name belonging to one of six race and ethnicity categories: Hispanic; non-Hispanic White; non-Hispanic Black or African American; non-Hispanic Asian/Pacific Islander; non-Hispanic American Indian and Alaska Native; and non-Hispanic Multiracial. These categories are consistent with 1997 Office of Management and Budget (OMB) definitions.[9,10] In total, the list provides 151,671 surnames, covering approximately 90% of the U.S. population. Word et al. (2008) provides a detailed description of how the census surname list was constructed and describes the routines used to standardize surnames appearing on the list.[11]

3.1.2 Geography

Information on the racial and ethnic composition of the U.S. population by geography comes from the Summary File 1 (SF1) from Census 2010, which provides counts of enumerated

---

[8] The data and documentation are available at: http://www.census.gov/genealogy/www/data/2000surnames/. The most recent census year for which the surname list exists is 2000. We will rely on more current data when it becomes available.

[9] This classification holds Hispanic as mutually exclusive from the race categories, with individuals identified as Hispanic belonging only to that category, regardless of racial background. The Census relies on self-identification of both race and ethnicity when determining race and ethnicity for these individuals, with an exception made for classification to the "Some Other Race" category. In Census 2000, some individuals identifying as "Some Other Race" also specified a Hispanic nationality (e.g., Salvadoran, Puerto Rican); in these instances, the Census identified the respondent as Hispanic. OMB definitions are available at: http://www.whitehouse.gov/omb/fedreg_1997standards.

[10] In the census surname data, the Census Bureau suppressed exact counts for race and ethnicity categories with 2-5 occurrences for a given name. Similarly to Elliott et al. (2009), in these cases we distribute the sum of the suppressed counts for each surname evenly across all categories with missing nonzero counts.

[11] Word, D.L., Coleman, C.D., Nunziata, R., Kominski, R., Demographic aspects of surnames from Census 2000. Available at: http://www.census.gov/genealogy/www/data/2000surnames/surnames.pdf.

individuals by race and ethnicity for various geographic area definitions, with census block serving as the highest level of disaggregation (the smallest geography).[12] In the decennial Census of the Population, the Census Bureau uses a classification scheme for race and ethnicity that differs slightly from the scheme used by OMB. Census treats Hispanic as an ethnicity and the other OMB categories as racial identities. However, Census does report population counts by race and ethnicity in a way that allows for the creation of race and ethnicity population totals that are consistent with the OMB definition.[13] Our method relies on race and ethnicity information for the adult (age 18 and over) population at the census block group, census tract, and 5-digit zip code levels, as discussed in the next section.[14,15]

3.2 Constructing the BISG probability

Constructing the BISG proxy for race and ethnicity for a given set of applicants requires place of residence (address) and name information for those applicants, the census surname list, and census demographic information by census block group, census tract, and 5-digit zip code. The process occurs in a number of steps:

1. Applicants' surnames are standardized and edited, including removing special characters and titles, such as JR and SR, and parsing compound names.

---

[12] The hierarchy of census geographic entities, from smallest to largest, is: block, block group, tract, county, state, division, region, and nation. Block group level information appears in Table P9 ("Hispanic or Latino, and Not Hispanic or Latino by Race") in the SF1. Table P11 in the SF1 provides similar counts for the restricted population of individuals 18 and over. The public can access these data in a variety of ways, including through the American FactFinder portal at: http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml.

[13] In the 2010 SF1, Census produced tabulations that report counts of Hispanics and non-Hispanics by race. These tabulations include a "Some Other Race" category. As in Elliott et al. (2009), we reallocate the "Some Other Race" counts to each of the remaining six race and ethnicity categories using an Iterative Proportional Fitting procedure to make geography based demographic categories consistent with those on the census surname list.

[14] Throughout this paper, we use 5-digit zip code, when referring to zip code demographics, as a synonym for ZIP Code Tabulation Areas (ZCTAs) as defined by the U.S. Census Bureau. More information on the construction of ZCTAs is available at: https://www.census.gov/geo/reference/zctas.html.

[15] From the SF1, we retain population counts for the contiguous U.S., Alaska, and Hawaii in order to ensure consistency with the population covered by the census surname list.

2. Standardized surnames are matched to the census surname list. For applicants with compound surnames, if the first word of the compound surname successfully matches to the surname data, it is used to calculate the surname based probability. If the first word does not match, the second word is then tried. For example, if an applicant's last name is Smith-Jones, the demographic information associated with Smith is used if Smith appears on the name list. If Smith does not appear on the name list, then the information associated with Jones is used if Jones is on the list.

3. For each name that matches the census surname list, the probability of belonging to a given racial or ethnic group (for each of the six race and ethnicity categories) is constructed. The probability is simply the proportion (or percentage) of individuals who identify as being a member of a given race or ethnicity for a given surname. For example, according to the census surname list, 73% of individuals with the surname Smith report being non-Hispanic White; thus, for any individual with the last name Smith, the surname-based probability of being non-Hispanic White is 73%. For applications with names that do not match the census surname list, a probability is not constructed. These records are excluded in subsequent analysis.[16] Given that approximately 10% of the U.S. population is not included on the census surname list, one would reasonably expect roughly a 10% reduction in the number of records in a proxied dataset due to non-matches to the census surname list.

4. Applicant address information is standardized in preparation for geocoding. Standardization includes basic checks such as removing non-numeric characters from zip codes, making sure zip codes with leading zeroes are accurately identified, and ensuring address information is in the correct format, for example, that house number, street, city, state, and zip code are appropriately parsed into separate fields.

5. Addresses are mapped into census geographic areas using a geocoding and mapping software application.[17] The geocoding application used by OR and SEFL in building the

[16] Elliott et al. (2009) retain records in their assessment data that do not appear on the surname list. To do so, they use the distribution of race and ethnicity appearing on the name list and the national population counts in the Census 2000 SF1 to characterize the unlisted population. OR and SEFL continue to evaluate the approach undertaken by Elliott et al. (2009) and may adopt a method for proxying the unlisted surname population in future updates to the proxy methodology.

[17] We currently use ArcGIS Version 10.1 with Street Map Premium 2011 Release 3 to geocode data when building the proxy. We may rely on updated releases as they become available or may move to different geocoding technology in the future. The BISG proxy methodology does not require the use of a specific geocoding technology.

proxy identifies the geographic precision to which an address is geocoded, and the precision of geocoding determines the precision of the demographic information relied upon.[18] For addresses that are geocoded to the latitude and longitude of an exact street address (often referred to as a "rooftop"), information on race and ethnicity for the adult population residing in the census block group containing the street address is used; if the census block group has zero population, information for the census tract is used. For addresses that are geocoded to street name, 9-digit zip code, and 5-digit zip code, the race and ethnicity information for the adult population residing in the 5-digit zip code is used. Addresses that cannot be geocoded or that can be geocoded only to a geographical area that is less precise than 5-digit zip code (for example, city or state) are excluded in subsequent analysis.

6. For geocoded addresses, the proportion (or percentage) of the U.S. adult population for each race and ethnicity residing in the geographic area containing the address or associated with the 5-digit zip code is calculated.

7. Bayes Theorem is used to update the surname-based probabilities constructed in Step 3 with the information on the concentration of the U.S. adult population constructed in Step 6 to create a probability—a value between, or equal to, 0 and 1—of assignment to each of the 6 race and ethnicity categories. These proxy probabilities can be used in statistical analysis aimed at identifying potential differences in lending outcomes.

Appendix A provides the mathematical formula associated with Step 7 and an example of the construction of the BISG proxy probabilities for an individual with the last name Smith residing in California. The statistical software code, written in Stata, and the publicly available census data files used to build the BISG proxy are available at: https://github.com/cfpb/proxy-methodology. Because OR and SEFL currently use ArcGIS to geocode address information when building the proxy, the geocoding of address information must occur before running the Stata code that builds the BISG proxy. The use of alternative geocoding applications may return slightly different geocoding results and, therefore, may yield different BISG probabilities than those generated using ArcGIS.

Steps 1 through 7 describe the general process currently undertaken by OR and SEFL to construct proxies for race and ethnicity for fair lending analysis. Unique features of a dataset

[18] The precision of the geocoding is driven by the availability of address information and the geocoding software application's assessment of the quality of address information provided.

under review, for example, the quality of surname data and the ability to match individuals to the census surname list, or the quality of address information and the ability to geocode to an acceptable level of precision, may lead to a modification of the general methodology, as appropriate.

4. Assessing the ability to predict race and ethnicity: an application to mortgage data

Elliott et al. (2009) demonstrate, using health plan enrollment data with reported race and ethnicity, that the BISG proxy methodology is more accurate than either the traditional surname-only or geography-only methodologies. In this section, we discuss a similar validation of the BISG proxy in the mortgage lending context.

To assess the performance of the BISG proxy in this context, the geography-only, surname-only, and BISG proxies for race and ethnicity were constructed for applicants appearing in a sample of mortgage loan applications in 2011 and 2012 for which address, name, and race and ethnicity were reported.[19,20] These data were provided to the CFPB by a number of lenders pursuant to the CFPB's supervisory authority. Applications with surnames that did not match the surname list

[19] The geography-only probability proxy is constructed in a manner that is similar to the construction of the surname-only proxy. For each geocoded address, the probability of belonging to a given racial or ethnic group (for each of the six race and ethnicity categories) is constructed. The probability is simply the proportion (or percentage) of individuals who identify as being a member of a given race or ethnicity who reside in the block group, census tract, or area corresponding to the 5-digit zip code, depending on the precision to which an applicant's address is geocoded.

[20] The reported race and ethnicity used in the assessment are derived from the HMDA reported race and ethnicity contained in the mortgage data sample. Ethnicity (Hispanic) and race—American Indian/Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White—are reported at the applicant level. For a given applicant, up to five races may be reported. The reported HMDA race and ethnicity are used to classify applicants in a manner consistent with the six mutually exclusive race and ethnicity categories defined by the Office of Management and Budget and used on the census surname list. Applications for which race or ethnicity information was not provided were omitted from the initial sample.

and with addresses that could not be geocoded to at least the 5-digit zip code were omitted from the analysis. Table 1 shows that for the initial sample of 216,798 mortgage applications, 26,363 applications—approximately 12% of the initial sample—were omitted from the analysis, resulting in a final sample of 190,435.

TABLE 1: MORTGAGE LOAN SAMPLE

	Not Geocoded	Geocoded
Surname did not match	8	26,297
Surname did match	58	190,435

For each applicant, three probabilities of assignment to each of the six race and ethnicity categories were constructed: a probability based on census race and ethnicity information associated with geography (geography-only); a probability based on census race and ethnicity information associated with surname (surname-only); and the BISG probability based on census race and ethnicity information associated with surname and geography (BISG). As previously discussed, the probabilities themselves may be used to proxy for race and ethnicity by assigning to each record a probability of belonging to a particular racial or ethnic group. These probabilities can be used to estimate the number of individuals by race and ethnicity and to identify potential disparities in outcomes through statistical analysis.

Assessing the accuracy of the proxy involves comparing a probability that can range between 0 and 1 (a continuous measure) to reported race and ethnicity classifications that, by definition, take on values of only 0 or 1 (a dichotomous measure). Accuracy can be evaluated in at least two ways: (1) by comparing the distribution of race and ethnicity across all applicants based on the proxy to the distribution based on reported characteristics and (2) by assessing how well the proxy is able to sort applicants into the reported race and ethnicity categories. The tendency for low values of the proxy to be associated with low incidence of individuals in a particular racial or ethnic group and for high values of the proxy to be associated with high incidence is measured by the correlation between the proxy and reported classification for a given race and ethnicity. Additional diagnostic measures, such as Area Under the Curve (AUC) statistics, reflect the extent to which a proxy probability accurately sorts individuals into target race and ethnicity and provides a statistical framework for assessing improvements in sorting attributable to the BISG proxy. Section 4 provides an evaluation of the use of the BISG probability proxy and

assesses performance relative to reported race and ethnicity, illustrating the merits of relying on the BISG probability proxy rather than on a proxy based solely on information associated with geography or surname alone.

4.1 Composition of lending by race and ethnicity

Table 2 provides the distribution of reported race and ethnicity (Reported) and the distributions based on the BISG, surname-only, and geography-only proxies. For the Reported row, the percentage in each cell is calculated as the sum of the reported number of individuals in each racial or ethnic group divided by the number of applicants in the sample (multiplied by 100). For the proxies, the percentage is simply the sum of the probabilities for each race and ethnicity divided by the number of applicants in the sample (multiplied by 100). For example, two individuals each with a 0.5 probability of being Black and a 0.5 probability of being White would contribute a count of 1 to both the Black and the White totals.

TABLE 2: DISTRUBUTION OF LOANS BY RACE AND ETHNICITY[21]

Classifier or Proxy	Hispanic	White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial
Reported	5.8%	82.9%	6.2%	4.5%	0.1%	0.4%
BISG	6.1%	79.7%	7.5%	5.0%	0.2%	1.4%
Surname-only	7.4%	75.4%	10.0%	4.9%	0.6%	1.7%
Geography-only	7.2%	78.6%	8.1%	4.8%	0.3%	1.0%

[21] In this table and in subsequent tables, we refer only to the race for a non-Hispanic race group. For instance, the "White" category refers to "Non-Hispanic White."

As the table indicates, all three proxies tend to approximate the reported population race and ethnicity. However, each also tends to underestimate the population of non-Hispanic Whites and overestimate the other race and ethnicity categories, which may reflect differences between the racial and ethnic composition of the census based populations used to construct the proxies and the racial and ethnic composition of individuals applying for mortgages.

Importantly, however, the BISG proxy comes closer to approximating the reported race and ethnicity than the traditional proxy methodologies, with the only exception being for Asian/Pacific Islanders and Multiracial. Though we see small absolute gains in accuracy from use of a BISG proxy for some groups relative to the traditional methods of proxying, these gains frequently represent a sizeable improvement in terms of relative performance. For example, the gap between reported race and estimated race for non-Hispanic Whites shrinks by 1.1% (from 82.9% – 78.6% = 4.3% to 82.9% – 79.7% = 3.2%) when moving from a geography-only to the BISG proxy. Given the initial gap of 4.3% this represents an almost 25% reduction in the difference between estimated and reported race. The gaps for non-Hispanic Black, non-Hispanic American Indian/Alaska Native, and Hispanic shrink in a similar manner. For non-Hispanic Asian/Pacific Islander, the gap between estimated and reported totals increases by 0.2% in absolute terms compared to the geography-only alternative and by 0.1% compared to the surname-only alternative. For the non-Hispanic Multiracial category, the BISG proxy does slightly better than the surname-only and slightly worse than the geography-only proxy in approximating the reported percentage.

4.2 Predicting race and ethnicity for applicants

4.2.1 Correlations between the proxy and reported race and ethnicity

Table 3 provides the correlations between reported race and ethnicity and the BISG, surname-only, and geography-only proxies.

TABLE 3: CORRELATIONS BETWEEN PROXY PROBABILITY AND REPORTED RACE AND ETHNICITY

Proxy	Hispanic	White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial
BISG	0.81	0.77	0.70	0.83	0.06	0.05
Surname-only	0.78	0.66	0.40	0.81	0.03	0.05
Geography-only	0.45	0.54	0.58	0.38	0.05	0.03

Correlation is a statistical measure of the relationship between different variables—in this case the race and ethnicity proxy and an applicant's reported race and ethnicity. Positive values indicate a positive correlation (as one variable increases in value, so does the other), negative values imply negative correlation (as one variable increases in value, the other decreases), and 0 indicates no statistical relationship. By definition, a correlation coefficient of 0 means that the proxy probability has no predictive power in explaining movement in the reported value, while a coefficient of 1 means that an increase in the proxy probability perfectly predicts increases in the reported values. Higher values of the correlation measure indicate a stronger ability to accurately sort individuals both into and out of a given race and ethnicity classification.

Correlations associated with the BISG proxy probabilities for Hispanic and non-Hispanic White, Black, and Asian/Pacific Islander are large and suggest strong positive co-movement with reported race and ethnicity. This means, for example, that the Hispanic proxy value is higher on average for individuals who are reported as Hispanic than for those who are not. For non-Hispanic American Indian/Alaska Native and the Multiracial classifications, correlations are positive but close to zero for all proxy methods, suggesting a low degree of power in predicting reported race and ethnicity for these two groups.

Looking across the rows in Table 3, correlations associated with the BISG are higher than those associated with the surname-only and geography-only proxies, notably for non-Hispanic Black and non-Hispanic White, reflecting the increase in the strength of the relationship between the proxy and reported characteristic from the integration of information associated with surname and geography in the BISG proxy. These results align closely with those found in Elliott et al.

(2009), which, as previously noted, assessed the BISG proxy using national health plan enrollment data.[22]

4.2.2 Area Under the Curve (AUC)

While correlations illustrate the overall extent of co-movement between the proxies and reported race and ethnicity, it is also important to assess the extent to which the proxy probabilities successfully sort individuals into each race and ethnicity.

A statistic that can be used to calculate this is called the Area Under the Curve (AUC), which represents the likelihood that the proxy will accurately sort individuals into a particular racial or ethnic group.[23] For example, if one randomly selects an individual who is reported as Hispanic and a second individual who is reported as non-Hispanic, the AUC represents the likelihood that the randomly selected individual reported as Hispanic has a higher proxy value of being Hispanic than the randomly selected individual reported as non-Hispanic. The AUC can be used to test the hypothesis that one proxy is more accurate than another at sorting individuals in order of likelihood of belonging to a given race and ethnicity. An AUC value of 1 (or 100%) reflects perfect sorting and classification, and a value of 0.5 (or 50%) suggests that the proxy is only as good as a random guess (e.g., a coin toss).

Table 4 provides the results of statistical comparisons of the geography-only, surname-only, and BISG probabilities. The AUC statistics associated with the BISG proxy for Hispanic and non-Hispanic White, Black, and Asian/Pacific Islander are large and exceed 90%. For instance, the AUC statistic associated with the BISG proxy for non-Hispanic Black is 0.9540, suggesting that 95% of the time, a randomly chosen individual reported as Black will have a higher BISG probability of being Black than a randomly chosen individual reported as non-Black.

---

[22] Table 4 of Elliott et al. (2009): Non-Hispanic White (0.76); Hispanic (0.82); Black (0.70); Asian/Pacific Islander (0.77); American Indian/Alaska Native (0.11); and Multiracial (0.02).

[23] The AUC is based on the Receiver Operating Characteristic (ROC) curve, which plots the tradeoff between the true positive rate and the false positive rate for a given proxy probability over the entire range of possible threshold values that could be used to classify individuals with certainty to the race and ethnicity being proxied. See Appendix B for more detail on the construction of the ROC curves and calculation of the AUC.

TABLE 4: LIKELIHOOD OF ASSIGNMENT OF HIGHER PROXY PROBABILITY FOR GROUP MEMBERSHIP GIVEN THAT BORROWER IS REPORTED AS MEMBER OF GROUP (AREA UNDER THE CURVE STATISTIC)

Proxy	Hispanic	White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial
BISG	0.9446	0.9430	0.9540	0.9723	0.6840	0.6846
Geography-only	0.8386	0.8389	0.8959	0.8359	0.6574	0.6015
Surname-only	0.9302	0.8968	0.8678	0.9651	0.5907	0.7075
p-value, H_0: BISG=Geo	<0.0001	<0.0001	<0.0001	<0.0001	0.0262	<0.0001
p-value, H_0: BISG=Name	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0289

For each of these four race and ethnicity categories, the AUC for the BISG proxy probability is statistically significantly larger than the AUC for the surname-only and geography-only probabilities, suggesting that, at or above the 99% level of statistical significance, the BISG more accurately sorts individuals than the traditional proxy methodologies.[24] The greatest improvements in the AUC are associated with the BISG proxy for non-Hispanic White and Black, as the AUC is considerably higher than the AUCs associated with the geography-only and surname-only proxies. For Hispanic and non-Hispanic Asian/Pacific Islander, this improvement is only marginal relative to the performance of the surname-only proxy. Performance for non-Hispanic American Indian/Alaska Native and Multiracial, while generally improved by the use of the BISG proxy probabilities, is weak overall regardless of proxy choice, with only an 18% improvement in sorting over a random guess. These results suggest that proxies based on census geography and surname data are not particularly powerful in their ability to sort individuals into these two race and ethnicity categories.

[24] The p-values for the tests of equivalence of the AUC statistics for the BISG and geography-only proxies and the BISG and surname-only proxies for each race and ethnicity appear in the last two rows of Table 4.

4.2.3 Classification over the range of proxy values

The BISG proxy's ability to sort individuals is made clear through an evaluation of the number of applicants falling within ranges of proxy probability values. For example, for 10% bands of the BISG proxy probability for Hispanics, Table 5 provides: the number of total applicants (column 1); the estimated number of Hispanic applicants based on the summation of the BISG probability (column 2); the number of reported Hispanic applicants (column 3); the number of reported non-Hispanic White applicants (column 4); and the number of reported other minority, non-Hispanic applicants (column 5). A few results are worth noting.

TABLE 5: CLASSIFICATION OVER RANGE OF BISG PROXY FOR HISPANIC

Hispanic BISG Proxy Probability Range	Total Applicants (1)	Estimated Hispanic (BISG) (2)	Reported Hispanic (3)	Reported White (4)	Reported Other Minority (5)
0% - 10%	176,116	1,129	1,677	153,974	20,465
10% - 20%	1,720	240	163	1,207	350
20% - 30%	653	163	130	414	109
30% - 40%	541	189	147	312	82
40% - 50%	557	251	226	261	70
50% - 60%	597	328	279	258	60
60% - 70%	802	522	455	263	84
70% - 80%	1,135	853	766	286	83
80% - 90%	1,788	1,529	1,347	347	94
90% - 100%	6,526	6,312	5,883	534	109
Total	190,435	11,516	11,073	157,856	21,506

**Estimated Hispanic (BISG) is calculated as the sum of the BISG probabilities for being Hispanic within the corresponding proxy probability range.*

First, the distribution of the BISG proxy probability is bimodal with concentrations of total applicants for low (e.g., 0%-20%) and high (e.g., 80%-100%) values of the proxy, which illustrates the sorting feature of the proxy. Reported Hispanic applicants are concentrated within high values of the proxy. For example, 65% ((1,347+5,883)/11,073) of reported Hispanic applicants (column 3) have BISG proxy probabilities greater than 80%; this concentration is mirrored by the estimated number of Hispanic applicants (column 2), 68% of whom have BISG proxy probabilities greater than 80% ((1,529+6,312)/11,516). While the BISG proxy may assign high values to some non-Hispanic applicants, 98% ((153,974+1,207)/157,856) of the reported non-Hispanic White and 97% ((20,465+350)/21,506) of the reported other non-Hispanic minority borrowers have Hispanic BISG proxy probabilities that are less than 20%.

Second, there are reported Hispanic applicants over the full range of values of the BISG proxy; this is also reflected by the estimated counts in column 2. For example, there are 597 applicants with BISG proxy values between 50% and 60%, of whom 279 are reported as being Hispanic, while the BISG proxy estimate of the number of Hispanic applicants in this range—calculated by summing probabilities for individuals within this probability range—is 328.

As suggested by Table 5 the BISG proxy tends to overestimate the number of Hispanic applicants for the mortgage pool under review. In the final row of column (3) we see that the total number of reported Hispanic applicants is 11,073. The estimated total number of Hispanic applicants—calculated as the sum of the BISG probabilities for Hispanic applicants—is 11,516 (column 2), which overestimates the number of Hispanic applicants by 4%. This overestimation may reflect, as discussed in Section 4.1, the use of demographic information based on the population at large to proxy the characteristics of mortgage applicants. According to the 2010 Census of Population, 14% of the U.S. adult population was Hispanic; 67% non-Hispanic White; 12% non-Hispanic Black; 5% Asian/Pacific Islander; and 1% American Indian/Alaska Native. According to the 2010 HMDA loan application data for all reporting mortgage originators, only 7% of applicants for home mortgages were Hispanic; 80% non-Hispanic White; 6% non-Hispanic Black; 6% Asian/Pacific Islander; and less than 1% American Indian/Alaska Native.[25] Mortgage borrowers tend to be disproportionately non-Hispanic White and, in particular, underrepresent Hispanic and non-Hispanic Blacks relative to the population of the U.S.

[25] The HMDA distributions for race and ethnicity are based only on applicant information for which race and ethnicity is reported and for applications that were originated, approved but not accepted, and denied by lenders.

OR and SEFL rely directly on the BISG probability in our fair lending related statistical analyses. In contrast, some practitioners rely on the use of a probability proxy and a threshold rule to classify individuals into race and ethnicity. When a threshold rule is used, individuals with proxy probabilities equal to and greater than a specific value, for example 80%, are considered to belong to a group with certainty, while all others are considered non-members with certainty. Consider two individuals who are assigned BISG probabilities of being non-Hispanic Black: individual A with 82% and individual B with 53%. The application of an 80% threshold rule for assignment would force individual A's probability to 100% and classify that individual as being Black and force individual B's probability to 0% and classify that individual as being non-Black.

The threshold rule removes the uncertainty about group membership at the cost of decreased statistical precision, with that precision deteriorating with decreases in the proxy's ability to create separation across races and ethnicity. In situations in which researchers can obtain clear separation between groups—for instance, situations for which the probabilities of assignment tend to be very close to 0 or 1—the consequences of using a threshold assignment rule, beyond simple measurement error, would be minor. However, when insufficient separation exists—for example, when there are a significant number of individuals with probabilities between 20% and 80% of belonging to a particular group—the use of thresholds can artificially bias, usually downward, estimates of the number of individuals belonging to particular racial and ethnic groups and potentially attenuate estimates of differences in outcomes between groups.

Table 5 makes clear the consequence of applying a threshold rule to the BISG proxy probability to force classification with certainty. If an 80% threshold rule is applied, the estimated number of Hispanic applicants is 8,314—the sum of all applicants in column (1) with a BISG probability equal to or greater than 80%—which underestimates the reported number of 11,073 Hispanic applicants by 25%. The underestimation is driven by the failure to count the large number of individuals in column (3) who are reported as being Hispanic in the mortgage sample but for whom the BISG probability of assignment is less than 80%.

It is worth noting that the application of an 80% threshold rule to classify individuals also yields false positives: individuals who are reported as being non-Hispanic but, nonetheless, are assigned BISG proxy probabilities of being Hispanic equal to or greater than 80%. For the mortgage pool under review, 881 applicants who are reported as being non-Hispanic White and 203 applicants who are reported as being some other minority would be classified as Hispanic by an 80% threshold rule. The false positive rate associated with these 1,084 observations is 0.6%, measured as the number of false positives (1,084) as a percentage of the total number of false positives plus the 178,278 true negative reported non-Hispanics with BISG probabilities

less than 80%. The false discovery rate for these same 1,084 observations is 13%, measured as the number of false positives (1,084) as a percentage of 8,314 applicants identified as Hispanic by the 80% threshold rule.

Classification and misclassification tables for the other five race and ethnicity categories appear in Appendix C.

5. Conclusion

Information on consumer race and ethnicity is generally not collected for non-mortgage credit products. However, information on consumer race and ethnicity is required to conduct fair lending analysis. Publicly available data characterizing the distribution of the population across race and ethnicity on the basis of geography and surname can be used to develop a proxy for race and ethnicity. Historically, practitioners have relied on proxies based on geography or surname only. A new approach proposed in the academic literature—the BISG method—combines geography- and surname-based information into a single proxy probability. In supervisory and enforcement contexts, OR and SEFL rely on a BISG proxy probability for race and ethnicity in fair lending analysis conducted for non-mortgage products.

This paper explains the construction of the BISG proxy currently employed by OR and SEFL and provides an assessment of the performance of the BISG method using a sample of mortgage applicants for whom race and ethnicity are reported. Our assessment demonstrates that the BISG proxy probability is more accurate than a geography-only or surname-only proxy in its ability to predict individual applicants' reported race and ethnicity and is generally more accurate than a geography-only or surname-only proxy at approximating the overall reported distribution of race and ethnicity. We also demonstrate that the direct use of the BISG probability does not introduce the sample attrition and significant underestimation of the number of individuals by race and ethnicity that occurs when commonly-relied-upon threshold values are used to classify individuals into race and ethnicity categories.

OR and SEFL do not require the use of or reliance on the specific proxy methodology put forth in this paper, but we are making available to the public the methodology, statistical software code, and our understanding of the performance of the methodology for a pool of mortgage applicants in an effort to foster transparency around our work. The methodology has evolved over time and will continue to evolve as enhancements are identified that improve accuracy and performance. Finally, the Bureau is committed to continuing our dialogue with other federal agencies, lenders, advocates, and researchers regarding the methodology.

6. Technical Appendix A: Constructing the BISG probability

For race and ethnicity, demographic information associated with surname and place of residence are combined to form a joint probability using the Bayesian updating methodology described in Elliott, et al. (2009). For an individual with surname *s* who resides in geographic area *g*:

1. Calculate the probability of belonging to race or ethnicity r (for each of the six race and ethnicity categories) for a given surname s. Call this probability p(r|s).
2. Calculate the proportion of the population of individuals in race or ethnicity r (for each of the six race and ethnicity categories) that lives in geographic area g. Call this proportion q(g|r).
3. Apply Bayes' Theorem to calculate the likelihood that an individual with surname s living in geographic area g belongs to race or ethnicity r. This is described by

$$\Pr(r|g,s) = \frac{p(r|s)q(g|r)}{\sum_{r\in R} p * q}$$

where *R* refers to the set of six OMB defined race and ethnicity categories. To maintain the statistical validity of the Bayesian updating process, one assumption is required: the probability of residing in a given geography, given one's race, is independent of one's surname. For example, the accuracy of the proxy would be impacted if Blacks with the last name Jones preferred to live in a certain neighborhood more than both Blacks in general and all people with the last name Jones.

Suppose we want to construct the BISG probabilities on the basis of surname and state of residence for an individual with the last name Smith who resides in California.[26] Table 6 provides the distribution across race and ethnicity for individuals in the U.S. with the last name Smith.[27] For individuals with the surname Smith, the probability of being non-Hispanic Black, based on surname alone, is simply the percentage of the Smith population that is non-Hispanic Black: 22.22%.

TABLE 6: DISTRIBUTION OF RACE AND ETHNICITY FOR INDIVIDUALS IN THE U.S. POPULATION WITH THE SURNAME SMITH

Race/Ethnicity	Distribution
Hispanic	1.56%
White	73.35%
Black	22.22%
Asian/Pacific Islander	0.40%
American Indian/Alaska Native	0.85%
Multiracial	1.63%

To update the probabilities of assignment to race and ethnicity, the percentage of the U.S. population residing in California by race and ethnicity is calculated. These percentages appear in Table 7.

[26] In the example, we choose to use state to make the example easy to understand. In practice, a finer level of geographic detail is used as discussed earlier.

[27] "Smith" is the most frequently occurring surname in the 2000 Decennial Census of the Population. There are 2,376,206 individuals in the 2000 Decennial Census of Population with the last name "Smith" according to the surname list (http://www.census.gov/genealogy/www/data/2000surnames/).

TABLE 7: POPULATION RESIDING IN CALIFORNIA AS A PERCENTAGE OF THE TOTAL U.S. POPULATION BY RACE AND ETHNICITY

Race/Ethnicity	U.S. Population	California Population	% of U.S. Population Residing in California
Hispanic	33,346,703	9,257,499	27.76%
White	157,444,597	12,461,055	7.91%
Black	27,464,591	1,655,298	6.03%
Asian/Pacific Islander	11,901,269	3,968,506	33.35%
American Indian/Alaska Native	1,609,046	126,421	7.86%
Multiracial	2,797,866	490,137	17.52%
Total	234,564,071	27,958,916	11.92%

Given the information provided in these two tables, we can now construct the probability that Smith's race is non-Hispanic Black, given surname and residence in California using Bayes' Theorem. The probability of being non-Hispanic Black for the surname Smith (22.22%) is multiplied by the percentage of the non-Hispanic Black population residing in California (6.03%) and then divided by the sum of the products of the surname-based probabilities and percentage of the population residing in California for all six of the race and ethnicity categories:

$$\frac{.2222 * .0603}{.7335 * .0791 + .0156 * 0.2776 + .2222 * .0603 + .0040 * .3335 + .0085 * .0786 + .0163 * .1752} \approx 16.61\%$$

This same calculation is performed for the remaining race and ethnicity categories. Table 8 provides the surname-only and updated BISG probabilities for all six race and ethnicity categories for individuals with the last name Smith residing in California.

TABLE 8: SURNAME-ONLY AND BISG PROBABILITIES FOR "SMITH" IN CALIFORNIA

Race/Ethnicity	Surname-only	BISG
Hispanic	1.56%	5.37%
White	73.35%	72.00%
Black	22.22%	16.61%
Asian and Pacific Islander	0.40%	1.65%
American Indian/Alaska Native	0.85%	0.83%
Multiracial	1.63%	3.54%

The impact of the adjustment of the surname based probabilities is readily apparent: the surname probability is weighted downward or upward depending on the degree of overrepresentation or underrepresentation of the population of a given race and ethnicity in California relative to the percentage of the U.S. population residing in California. For example, just under 12% of the U.S. population resides in California but nearly 28% of Hispanics in the U.S. reside in California. Knowing that Smith resides in California and that California is more heavily Hispanic than the nation as a whole leads to an increase in the probability that Smith is Hispanic compared to the probability calculated based on surname information alone.

7. Technical Appendix B: Receiver Operating Characteristics and Area Under the Curve

One way to characterize the proxy's ability to sort individuals into race and ethnicity is to plot the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed by applying a threshold rule for classification to each race and ethnicity, where probabilities above the threshold yield classification to a given race and ethnicity and those below do not, and then plotting the relationship between the false positive rate and the true positive rate over the range of possible threshold values.

Figures 1 through 6 show the ROC curves for the geography-only, name-only, and BISG probabilities by race and ethnicity. In each plot, the true positive rate is measured on the y-axis and the false positive rate is measured on the x-axis.[28] The slope of the ROC curve represents the tradeoff between identifying true positives at the expense of increasing false positives over the range of possible threshold values. The ROC curve for a perfect proxy—one that could classify individuals into and out of a given race and ethnicity with no misclassification—moves along the edges of the figure from (0,0) to (0,1) to (1,1). The closer that the ROC curve is to the left and upper edge of the plot area, the better the proxy is at correctly classifying individuals. A proxy

[28] The true positive rate is defined as the ratio of the number of applicants correctly classified into a reported race and ethnicity by a given threshold divided by the total number applicants reporting the race and ethnicity; the false positive rate is defined as the ratio of applicants incorrectly classified into a reported race and ethnicity by a given threshold divided by the total number of applicants not reporting the race and ethnicity.

that provides no useful information instead moves along the 45-degree line that runs through the middle of the figure. Movement along this line implies that a proxy measure has no ability to meaningfully identify more true members of a group without simultaneously identifying a similar proportion of non-members.

The graphs demonstrate that for Hispanic and non-Hispanic White, Black, and Asian/Pacific Islander, the BISG proxy is generally associated with a higher ratio of true positives to false positives across all possible threshold values, as shown by the general tendency for BISG's ROC curve to be located to the left and above of the ROC curves for the surname-only and geography-only proxies. The BISG proxy's overall ability to improve sorting, relative to the surname-only or geography-only proxy, is especially notable for non-Hispanic Whites and Blacks. The AUC statistic discussed in Section 4.2.2 simply represents the area beneath the ROC curve and above the x-axis.

FIGURE 1: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR NON-HISPANIC WHITE

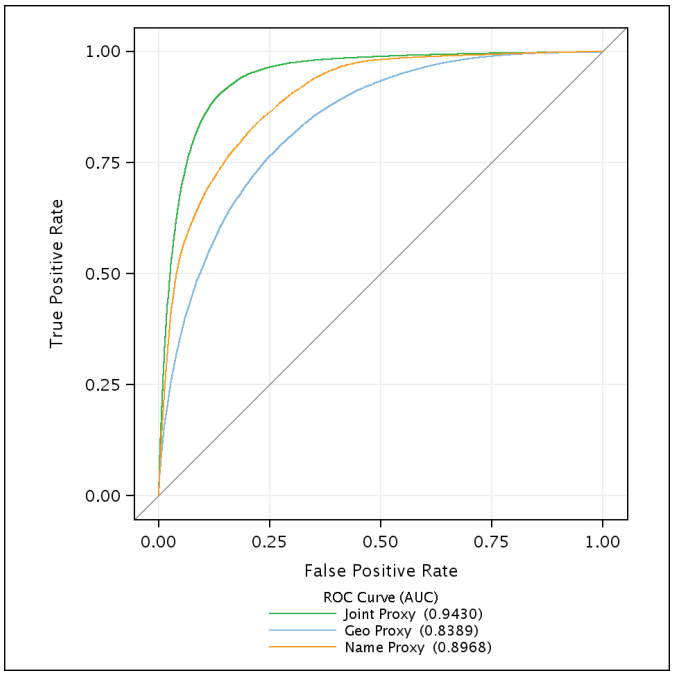


FIGURE 2: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR NON-HISPANIC BLACK

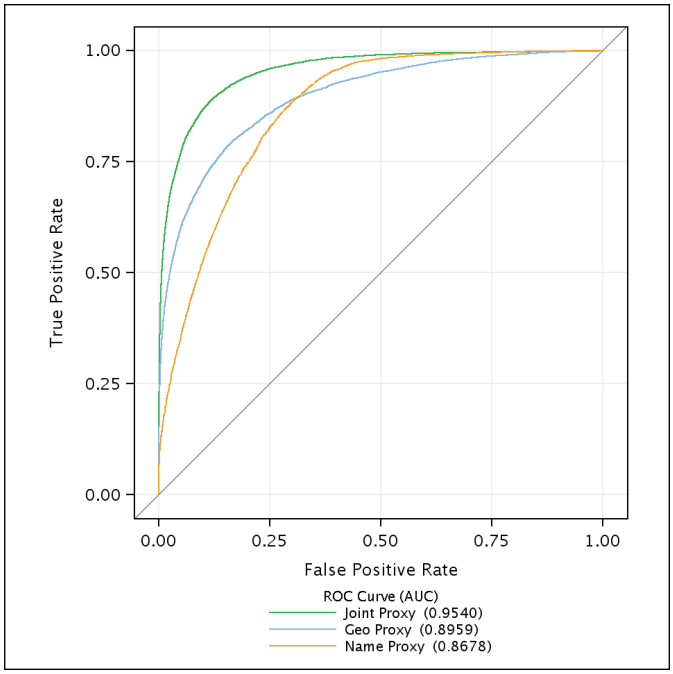


FIGURE 3: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR HISPANIC

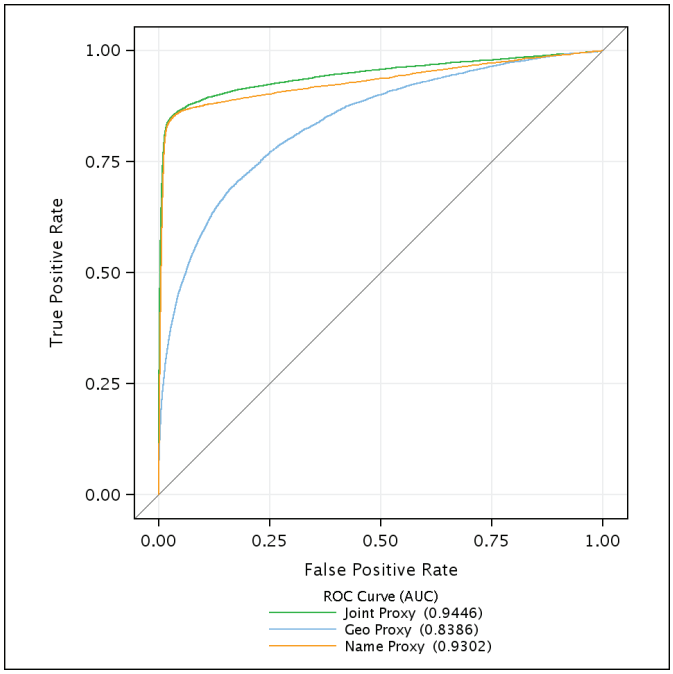


FIGURE 4: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR NON-HISPANIC ASIAN/PACIFIC

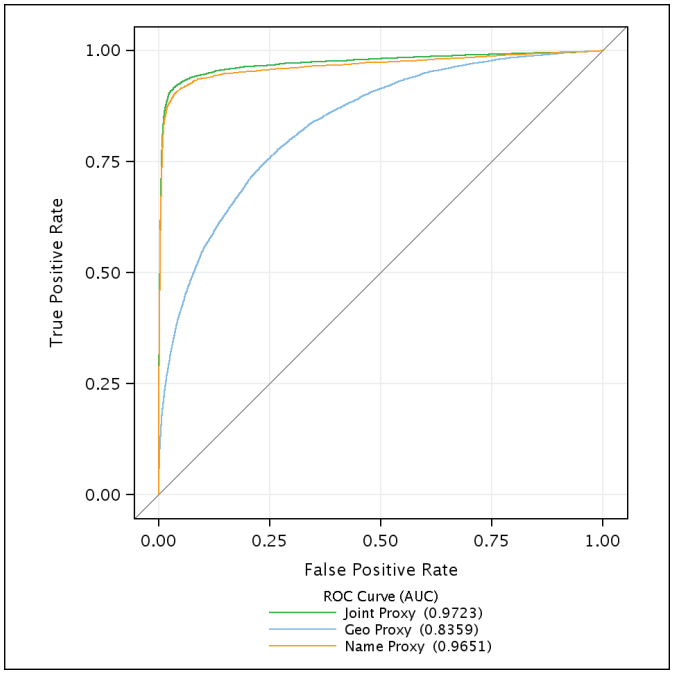


FIGURE 5: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR NON-HISPANIC NATIVE

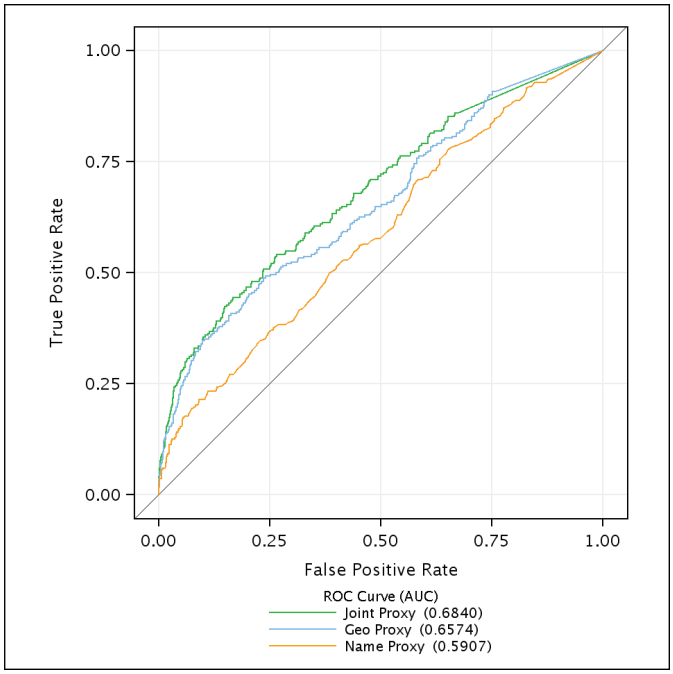
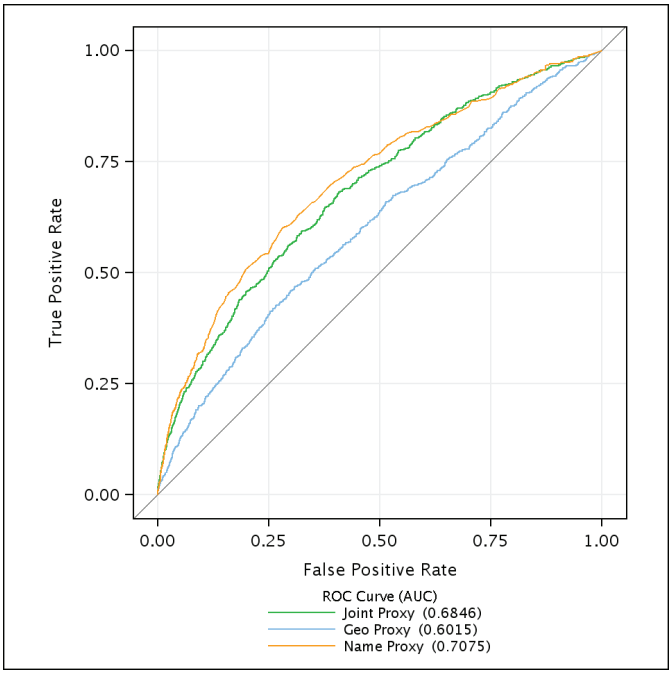


FIGURE 6: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR NON-HISPANIC MULTIRACIAL



8. Technical Appendix C: Additional tables

TABLE 9: CLASSIFICATION OVER RANGES OF BISG PROXY FOR NON-HISPANIC WHITE

White BISG Proxy Probability Range	Total Applicants (1)	Estimated White (BISG) (2)	Reported White (3)	Reported Minority (4)
0% - 10%	20,108	506	2,114	17,994
10% - 20%	3,995	582	937	3,058
20% - 30%	2,738	680	962	1,776
30% - 40%	2,483	867	1,206	1,277
40% - 50%	2,748	1,240	1,596	1,152
50% - 60%	3,346	1,847	2,196	1,150
60% - 70%	4,480	2,927	3,477	1,003
70% - 80%	7,105	5,363	5,851	1,254
80% - 90%	15,620	13,409	14,201	1,419
90% - 100%	127,812	124,411	125,316	2,496
Total	190,435	151,832	157,856	32,579

TABLE 10: CLASSIFICATION OVER RANGES OF BISG PROXY FOR NON-HISPANIC BLACK

Black BISG Proxy Probability Range	Total Applicants (1)	Estimated Black (BISG) (2)	Reported Black (3)	Reported White (4)	Reported Other Minority (5)
0% - 10%	160,733	1,859	1,466	139,684	19,583
10% - 20%	9,742	1,387	941	8,403	398
20% - 30%	4,916	1,207	906	3,814	196
30% - 40%	3,101	1,072	726	2,242	133
40% - 50%	2,229	997	738	1,408	83
50% - 60%	1,680	922	736	877	67
60% - 70%	1,417	920	765	596	56
70% - 80%	1,407	1,057	963	391	53
80% - 90%	1,517	1,293	1,222	241	54
90% - 100%	3,693	3,548	3,408	200	85
Total	190,435	14,262	11,871	157,856	20,708

TABLE 11: CLASSIFICATION OVER RANGES OF BISG PROXY FOR NON-HISPANIC ASIAN/PACIFIC ISLANDER

Asian/ Pacific Islander BISG Proxy Probability Range	Total Applicants (1)	Estimated Asian and Pacific Islander (BISG) (2)	Reported Asian and Pacific Islander (3)	Reported White (4)	Reported Other Minority (5)
0% - 10%	178,533	867	861	154,872	22,800
10% - 20%	1,536	216	234	890	412
20% - 30%	657	160	147	366	144
30% - 40%	492	170	157	247	88
40% - 50%	385	174	145	176	64
50% - 60%	361	199	168	139	54
60% - 70%	411	267	223	156	32
70% - 80%	649	488	421	180	48
80% - 90%	1,268	1,085	923	270	75
90% - 100%	6,143	5,941	5,367	560	216
Total	190,435	9,567	8,646	157,856	23,933

TABLE 12: CLASSIFICATION OVER RANGES OF BISG PROXY FOR NON-HISPANIC AMERICAN INDIAN/ALASKA NATIVE

American Indian/Alaska Native BISG Proxy Probability Range	Total Applicants (1)	Estimated American Indian/Alaska Native (BISG) (2)	Reported American Indian/Alaska Native (3)	Reported White (4)	Reported Other Minority (5)
0% - 10%	190,212	377	238	157,680	32,294
10% - 20%	137	19	3	106	28
20% - 30%	38	9	2	30	6
30% - 40%	12	4	1	9	2
40% - 50%	15	7	1	13	1
50% - 60%	6	3	0	6	0
60% - 70%	5	3	1	4	0
70% - 80%	4	3	1	3	0
80% - 90%	1	1	1	0	0
90% - 100%	5	5	0	5	0
Total	190,435	431	248	157,856	32,331

TABLE 13: CLASSIFICATION OVER RANGES OF BISG PROXY PROBABILITIES FOR NON-HISPANIC MULTIRACIAL

Multiracial BISG Proxy Probability Range	Total Applicants (1)	Estimated Multiracial (BISG) (2)	Reported Multiracial (3)	Reported White (4)	Reported Other Minority (5)
0% - 10%	187,964	2,102	682	156,439	30,843
10% - 20%	1,615	224	34	937	644
20% - 30%	443	107	8	255	180
30% - 40%	199	68	5	115	79
40% - 50%	113	50	9	47	57
50% - 60%	56	31	3	34	19
60% - 70%	33	21	0	18	15
70% - 80%	9	7	0	8	1
80% - 90%	3	2	0	3	0
90% - 100%	0	0	0	0	0
Total	190,435	2,612	741	157,856	31,838

HSR HEALTH SERVICES RESEARCH
Impacting Health Practice and Policy Through State-of-the-Art Research and Thinking

Health Serv Res. 2008 Oct; 43(5 Pt 1): 1722–1736.
doi: 10.1111/j.1475-6773.2008.00854.x

PMCID: PMC2653886
PMID: 18479410

A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity

MarcN Elliott, Allen Fremont, PeterA Morrison, Philip Pantoja, and Nicole Lurie

Address correspondence to Marc N. Elliott, Ph.D., RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; e-mail: elliott@rand.org. Allen Fremont, M.D., Ph.D., and Philip Pantoja, M.A., are with the RAND Corporation, Santa Monica, CA. Peter A. Morrison, Ph.D., is with the RAND Corporation, Nantucket, MA. Nicole Lurie, MD, is with the RAND Corporation, Arlington, VA.

Abstract

Objective

To efficiently estimate race/ethnicity using administrative records to facilitate health care organizations' efforts to address disparities when self-reported race/ethnicity data are unavailable.

Data Source

Surname, geocoded residential address, and self-reported race/ethnicity from 1,973,362 enrollees of a national health plan.

Study Design

We compare the accuracy of a Bayesian approach to combining surname and geocoded information to estimate race/ethnicity to two other indirect methods: a non-Bayesian method that combines surname and geocoded information and geocoded information alone. We assess accuracy with respect to estimating (1) individual race/ethnicity and (2) overall racial/ethnic prevalence in a population.

Principal Findings

The Bayesian approach was 74 percent more efficient than geocoding alone in estimating individual race/ethnicity and 56 percent more efficient in estimating the prevalence of racial/ethnic groups, outperforming the non-Bayesian hybrid on both measures. The non-Bayesian hybrid was more efficient than geocoding alone in estimating individual race/ethnicity but less efficient with respect to prevalence ($p<.05$ for all differences).

Conclusions

The Bayesian Surname and Geocoding (BSG) method presented here efficiently integrates administrative data, substantially improving upon what is possible with a single source or from other hybrid methods; it offers a powerful tool that can help health care organizations address disparities until

self-reported race/ethnicity data are available.

Keywords: Bayes's theorem, health disparities, health plans, race, surname

---

Efforts to measure, monitor, and address racial/ethnic disparities in health care have been limited by the paucity of data regarding the race/ethnicity of users of the health care system. Indeed, until recently, many viewed the collection of such data as illegal (Fremont and Lurie 2004). One result is that the preponderance of studies on racial/ethnic differences in quality of care and patient outcomes has been limited to patients enrolled in Medicare or Medicaid. Several reports from the Institute of Medicine and the National Academy of Sciences recommend universal collection of self-reported data regarding race, ethnicity, and socioeconomic status as a first step toward addressing disparities (Institute of Medicine 2002; National Research Council 2004). While self-reported data are widely considered to be the gold standard, absent a mandate to do so, collection of such data will be slow and inconsistent.

Several efforts to collect and use such data are underway. For example, the Health Research and Educational Trust, an independent research affiliate of the American Hospital Association, has developed a toolkit for and is assisting a growing number of hospitals with collection of racial, ethnic, and language data. Similarly, a group of hospitals funded by the Robert Wood Johnson Foundation to address disparities in cardiovascular care have committed to collecting race/ethnicity data and monitoring quality of care for different racial/ethnic groups. State policy has also moved toward collecting racial/ethnic data. For example, as part of the Massachusetts health care reform legislation, collection of race/ethnicity data from all hospitalized patients is required by law (Boston Public Health Commission 2006). In California, SB 853 and related regulations require HMO plans to collect race, ethnicity, and language information (California State Senate 2007). Finally, several of the plans participating in the National Health Plan Collaborative to Improve Quality and Eliminate Disparities have begun voluntary collection of self-reported data on the race/ethnicity of their enrollees (National Health Plan Collaborative 2006). Aetna has the most experience in doing so, but even with a mandate from their CEO and significant investment of resources over the past 4 years, the plan has been able to obtain data on only one-third of their enrollees thus far. Although a few smaller regional plans that followed Aetna's lead have obtained a similar proportion of self-reported data in less time, completing the process will likely take several more years.

Surname and Geocoding Approaches

Because the process of obtaining self-reported race/ethnicity data can take years to complete, investigators have developed methods of estimating race/ethnicity indirectly from other sources. Two such methods are geocoding and surname analysis. Geocoding uses an individual's address to link individuals to census data about the geographic areas where they live. For example, knowing that a person lives in a Census Block Group (a small neighborhood of approximately 1,000 residents) where 90 percent of the residents are African American provides useful information for estimating that person's race.

Surname analysis infers race/ethnicity from surnames (last names). Insofar as a particular surname belongs almost exclusively to a particular group (as defined by race, ethnicity, or national origin), it is possible to identify its holder's probable membership in the group by using well-formulated surname dictionaries. Such dictionaries now exist for identifying Hispanics and various Asian nationalities (Perkins 1993; Abrahamse, Morrison, and Bolton 1994; Kestenbaum et al. 2000; Lauderdale and Kestenbaum 2000; Falkenstein 2002). Separate surname lists have been generated for Chinese, Indian, Japanese, Korean, Filipino, and Vietnamese Americans. Experimental dictionaries for identifying Arab Americans are under development (Morrison et al. 2003). Both surname analysis and geocoding have

recognized limitations—the former has almost no ability to distinguish blacks from non-Hispanic whites whereas the latter has little ability to identify Hispanics or Asians. Although these limitations have been partially overcome by combining the two approaches, the accuracy of prior combined approaches varies widely by geographic area, depending on the prevalence and degree of segregation of racial/ethnic groups (Fremont et al. 2005; Fiscella and Fremont 2006).

A New Hybrid Approach

To further address limitations of current indirect estimation approaches, we developed a new hybrid approach using Bayes's theorem. Bayes's theorem is commonly applied to medical diagnostic testing; in the context of evaluating diagnostic tests, the probability of a given individual having a disease depends both upon (1) an individual's prior probability of having the disease (usually determined from a base rate appropriate to the individual's risk group) and (2) the result of a diagnostic test. Bayes's Theorem updates prior probabilities with test results by considering the *sensitivity, Se* (probability of a positive test result for a positive individual), and *specificity, Sp* (probability of a negative test result for a negative individual), of the diagnostic test to produce an updated (posterior) probability, called the *positive predictive value, PPV*, that efficiently incorporates both sources of information using the formula:

$$PPV = P \times Se/(P \times Se + (1 - P) \times (1 - Sp))$$

Here, we extend the approach from the two-category prior probability that characterizes baseline disease prevalence rates and treat the racial/ethnic distribution of where an individual lives as a four-category prior, the categories being Hispanic, African American, Asian, and non-Hispanic white or other. Our “baseline prevalence” is based on the racial/ethnic composition of the Census Block Group to which the residence of the individual was geocoded. We treat the combined results of the Census Bureau Spanish Surname List and the Lauderdale–Kestenbaum Asian Surname List as another diagnostic test with three possible outcomes (surname appears on Asian list regardless of appearance on Hispanic list, surname appears on Spanish but not Asian list, surname appears on neither surname list).

Using a more general form of Bayes's Theorem, we then use the surname lists to update the prior probabilities of membership in each of the four race/ethnic categories with the surname list results to produce efficient, updated posterior probabilities of membership in the four groups. The extent of this updating increases with the sensitivity and specificity of the surname lists for the population in question. We refer to this new hybrid method as the *Bayesian Surname and Geocoding* method (BSG) to note that it uses a Bayesian approach to combine surname and geocoded information. These probabilities, in turn, can be used to estimate racial/ethnic composition. Though not the focus of the current validation analyses reported here, the estimates can also be used to identify possible disparities in health care or in health outcomes by race/ethnicity.

We compare the accuracy of BSG in estimating race/ethnicity to two other approaches, in all instances evaluating performance against a gold standard of self-report. The first alternative approach is a previous algorithm for combining the two information sources (Fremont et al. 2005; Fiscella and Fremont 2006) that we will here call the *Categorical Surname and Geocoding* approach (CSG) in order to note that it combines surname and geocoded information in a categorical fashion, described below. The second approach to which we compare BSG is one based solely on the geocoded racial/ethnic composition of the Census Block Group where each member lives. We call this final strategy the *Geocoding Only* (GO) approach. These three approaches are summarized in Table 1.

Table 1

Summary of Three Methods Compared

Method	*Needs/Uses Surnames*	*Needs/Uses Addresses*	*How It Works*	*Output*
BSG	Yes	Yes	Uses surname lists to update geocoded information and derive posterior probabilities	Probability
GO	No	Yes	Uses geocoded probabilities directly	Probability
CSG	Yes	Yes	Classifies Asians and Hispanics using surname lists; classifies others according to prevalence of blacks in block group	Classification

Methods

Data

We used national enrollment data from Aetna, a large national health plan. The data set consists of self-reported race/ethnicity (as a "gold-standard" used for validation), surname, geocoded address of residence (Census 2000 Block Group level, using the SF1 file), and gender for all 1,973,362 enrollees who voluntarily provided this information to the plan for quality monitoring and improvement purposes. While voluntarily reported race/ethnicity was predominantly non-Hispanic white or other (78.1 percent), the data set included a reasonable distribution of Hispanics (8.9 percent), blacks (8.0 percent), and Asians (5.0 percent); 51.2 percent (1,010,043) were female. Data disclosed to RAND were done so in compliance with HIPAA regulations.

Implementation of the BSG

The Appendix S1 describes the implementation of the BSG algorithm in detail. If the BSG produced classifications instead of probabilities, we could describe its performance in terms of the sensitivity and specificity of the BSG. Instead, we use alternative measures described below. The sensitivities and specificities of the *surname lists* do play a role with BSG, however. They are *inputs* or tuning parameters that determine how the geocoded and surname data are combined to produce posterior probabilities, as detailed in the Appendix S1 (the greater the sensitivity and specificity, the more the surname results change the probabilities derived from geocoding). Thus these surname list sensitivities and specificities are not directly evaluative of performance in this context, but are primarily intermediate parameters.

As applied to the primary data set, the sensitivity of the Spanish and Asian surname lists themselves were calculated at 80.4 and 51.5 percent, respectively. The specificities are 97.8 and 99.6 percent, respectively. These sensitivities and specificities are characteristics of the surname lists, not of the BSG. Table S1 describes the probability of members of a given group appearing on each surname list or neither given these sensitivities and specificities. For example, Asians will appear on the Asian list 51.5 percent of the time (irrespective of appearance on the Spanish list), on the Spanish list but not the Asian list 1.1 percent of the time, and on neither list 47.4 percent of the time at these levels of sensitivity and

specificity under the assumptions stated earlier.

Because we find higher sensitivity for males than females (83.1 versus 77.8 percent on the Spanish Surname List; 52.7 versus 50.2 percent on the Asian Surname List, $p<.05$ for each) and slightly higher specificity for males than females for the Spanish Surname List (98.0 versus 97.5 percent, $p<.05$) that are presumably related to retention of surnames after marriage, the BSG uses gender-specific sensitivities and specificities. Thus, for example, a male who appears on the Spanish surname list in a given block group receives a slightly higher posterior probability of being Hispanic than a female who appears on that same list from the same block group because the surname list is known to be more accurate for males than females. The Appendix S1 provides additional examples of how the BSG generates posterior probabilities as well as other details of its implementation.

Other Algorithms Used for Comparison with the BSG

The second method, GO, simply uses the racial/ethnic prevalences from Census Block Groups as probabilities. Surname lists provide no means by which to distinguish blacks from non-Hispanic whites, so do not permit estimates of disparities between these two groups. For this reason, a "surname only" approach is not considered.

Instead, we consider a previously described alternative combination of geocoding and surname information, the CSG (Fiscella and Fremont 2006). CSG categorizes individuals through a series of steps. It (1) labels a person Hispanic if their name appears on the Spanish surname list; if not, it (2) labels a person Asian if the name appears on the Asian surname list; if neither of these applies, geocoded race/ethnic information is used to adjudicate classifications among the remaining individuals into black or non-Hispanic white categories. In particular, (3) if an individual not appearing on either surname list resides in a block group that is at least 66 percent black, they are classified as black; (4) otherwise they are classified as non-Hispanic white. In an application using Medicare enrollees in a national health plan, this algorithm produced estimates of racial/ethnic heath disparities that were similar to those obtained with self-reported race-ethnicity (Fremont et al. 2005; Fiscella and Fremont 2006).

Outputs of BSG, CSG, and GO: Classifications versus Probabilities

CSG discretely classifies each plan member into one of four racial/ethnic categories, whereas BSG and GO produce probabilities of membership in each of these four groups. As an illustration, consider a hypothetical Bob Jones living in a Census Block Group that was 67 percent white/other, 11 percent black, 11 percent Hispanic, and 11 percent Asian. CSG would note that "Jones" was on neither surname list and that his block group was <66 percent black and would therefore classify Mr. Jones as white/other. GO would simply use these four prevalences as probabilities and estimate that Mr. Jones had a 67 percent chance of being white/other and an 11 percent chance of being a member of each of the other three groups. As illustrated in Table 2, BSG would note that "Jones" was on neither surname list and integrate that information with the sensitivities and specificities of those lists, as well as the racial/ethnic composition of his block group to estimate that Mr. Jones has a 78.7 percent chance of being white/other, a 12.9 percent chance of being black, a 6.1 percent chance of being Asian, and a 2.2 percent chance of being Hispanic. Note that being on neither surname list makes white/other and black more likely than they were before surnames were considered, and that the probability of being Hispanic falls more than the probability of being Asian (because the Spanish surname list has greater sensitivity than the Asian list). Additional examples appear in Table S3.

Table 2

Illustration of BSG Posterior Probabilities of the Race/Ethnicity of a Male Individual Living in a Census Block Group That Was 67 Percent White/Other and 11 Percent Each Asian, Hispanic, and Black

	BSG Posterior Probability of Race/Ethnicity			
Surname	***Asian***	***Hispanic***	***Black***	***White/Other***
Wang	0.937	0.008	0.008	0.048
Martinez	0.010	0.845	0.021	0.125
Jones	0.061	0.022	0.129	0.787

One can estimate prevalences, means, and disparities by race/ethnicity by working directly with probabilities, without ever producing individual classifications. For example, if one's goal were a prevalence estimate, averaging probabilities is more accurate than classifying and rounding before summing (McCaffrey and Elliott forthcoming). For example, in an area with 10 people who had a 57 percent chance of being white and a 43 percent chance of being black and another 10 people with a 69 percent chance of being white and a 31 percent chance of being black, racial/ethnic prevalences would be more accurately estimated as 63 percent white and 37 percent black (averaging probabilities) than as 100 percent white (classifying each person into the group that was most likely for them). Please see Table S4 for additional examples. Similarly, if the goal is to compare racial/ethnic groups in terms of a clinical process measure, such as adherence to diabetes care recommendations as measured by administrative records, one need not classify individuals into discrete categories. Instead, one can enter an individual's probabilities of membership in each of several racial/ethnic groups (omitting one as a reference group) as predictors in a linear or logistic regression and the coefficients will be unbiased estimates of the difference of each racial/ethnic group from the reference racial/ethnic group in the outcome. Moreover, McCaffrey and Elliott show that such direct use of these probabilities, while less accurate than truly knowing race/ethnicity with certainty for each individual, is more accurate and efficient than using categorical classifications based on these probabilities. In each of these instances, categorizing continuous probabilities into discrete classifications is an unnecessary step that discards substantial information by ignoring distinctions in probabilities. While there may be some instances in which one must make a discrete decision for specific individuals (e.g., whether to mail Spanish-language materials to specific addresses), direct use of probabilities will be more efficient for aggregate statistical inferences, including the comparison of racial/ethnic groups.

If we were only examining CSG, we could describe its accuracy of classification in terms of sensitivity, specificity, and positive predictive value. Because we are comparing both classification-based and probability-based methods, we employ different performance measures.

Evaluation

We compare BSG, CSG, and GO in terms of how closely the estimates of race/ethnicity that they produce match those derived from self-reported race/ethnicity for the same individuals. We develop two

performance metrics applicable to all three approaches (BSG, CSG, and GO). We then compare the relative efficiency of the three methods according to these two metrics. The first metric assesses accuracy in matching the four-category distribution of self-reported racial/ethnic prevalence in a population. The second metric assesses the accuracy of predicting individual race/ethnicity—the extent to which those who self-report a given race/ethnicity are assigned higher probabilities of that race/ethnicity (or are more likely to be classified as that race/ethnicity). The two measures are complementary in that the first detects systematic errors in four-category classifications (e.g., a method is overly likely to classify someone as white and insufficiently likely to classify someone as black), and the second measure detects unsystematic errors (e.g., a method doesn't overestimate or underestimate any group in aggregate, but is just not very accurate in predicting the race/ethnicity of specific individuals).

Performance Metric for Predicting Racial/Ethnic Prevalence

For each of the three methods, we report the prevalence estimates derived for each of four racial/ethnic groups and compare these with self-reported proportions. In order to summarize the accuracy across these four categories, we compute the average error of the four categorical racial/ethnic prevalences estimates, weighted by their true (self-reported proportions). Ratios of average squared errors can be used to measure the *relative efficiency* of two methods in estimating prevalences. To say that method one has a relative efficiency of 3.0 relative to method two means that the accuracy of method one using a given sample size is the same as what would be obtained with three times the sample size using method two.

Performance Metric for Predicting Individuals' Race/Ethnicity

The Brier score (Brier 1950) is the mean squared deviation of a prediction from the true corresponding dichotomous outcome. The Murphy decomposition of the Brier score (Yates 1982) distinguishes (a) uncontrollable variation due to the prevalence of the outcome from (b) the extent to which predictions correlate with the dichotomous outcome. We use this correlation (b) as our measure of performance in predicting individual race/ethnicity. This metric rescales predictive performance to a (0, 1) scale regardless of prevalence.

In particular, we use the correlation of the dichotomous or probabilistic prediction with a dichotomous indicator of true self-reported race-ethnicity for each of four racial/ethnic groups. Whether a method produces classifications or probabilities, it is a comparable measure of the accuracy with which individual race/ethnicity is predicted. Estimates for the four racial/ethnic measures are not independent, but are negatively correlated. To summarize performance across all four racial/ethnic categories, we also calculate an average correlation, weighted by prevalence, for each method. By comparing ratios of squared correlations, we can compare the relative efficiency of methods in predicting individual race/ethnicity.

Results

Predicting Racial/Ethnic Prevalences: Comparing BSG, CSG, and GO

Table 3 displays the overall proportions of self-reported race/ethnic data falling into the four categories, along with estimates derived from each of the three methods using the primary data set. The average deviation from self-report is also displayed for each method. When comparing methods, it may be noted that the sampling error in assessing accuracy in prevalence is sufficiently small that all differences of 0.1 percent or more are statistically significant. GO substantially overestimates the prevalence of Hispanics, moderately overestimates the prevalence of blacks, and moderately underestimates the

prevalence of Asians ($p<.05$ for each). CSG is very accurate for Hispanics, but it underestimates the prevalence of Asians by nearly a factor of two and underestimates the prevalence of blacks by nearly a factor of three ($p<.05$ for both). These patterns result in overestimating the proportion of plan members who are white.

Table 3

Comparing Overall Racial/Ethnic Prevalence Estimates to Self-Report Estimates (n=1,973,362)

	Estimated Percentage in Each Group				
	Hispanic	***Asian***	***Black***	***White/Other***	***Weighted Average Overall Deviation from Self-Report***
SELF-REPORT	**8.9**	**5.0**	**8.0**	**78.1**	(0)
BSG	10.0	4.5	9.1	76.4	1.6%
GO	10.8	4.2	9.0	76.0	2.0%
CSG	9.2	2.9	3.0	84.9	6.2%

Ninety-five percent margins of sampling error are <0.1% for a single prevalence estimate, a difference in prevalences estimates across methods.

BSG is the most accurate overall, with a weighted average prevalence error (deviation from self-reported) of 1.6 percent, followed by 2.0 percent for GO and 6.2 percent for CSG ($p<.05$ for all pairwise comparisons). BSG moderately overestimates Hispanic and black prevalence, while underestimating whites and Asians somewhat ($p<.05$ for each). BSG is 56 percent more efficient than geocoding alone in prevalence estimates, whereas CSG is less efficient for this purpose than geocoding alone.

Predicting Individual Race/Ethnicity: Comparing BSG, CSG, and GO

Table 4 displays the correlation with self-reported race/ethnicity for each of the three methods and four race/ethnic groups in the primary data set. All reported correlations are statistically significant and differ across methods at $p<.05$. BSG predictions correlate with individual indicators of race/ethnicity at 0.61 to 0.79, with a weighted average correlation of 0.70.

Table 4

Correlation of Individual Predicted Race/Ethnicity with Self-Reported Race/Ethnicity (*n*=1,973,362)

	Correlation with Self-Reported Race/Ethnicity				
	Hispanic	***Asian***	***Black***	***White/Other***	***Weighted Average***
BSG	0.79	0.67	0.61	0.70	0.70
GO	0.49	0.34	0.57	0.55	0.53
CSG	0.77	0.65	0.48	0.63	0.63

All differences in correlations by methods are significant at $p<.05$.

CSG is the next best by this measure (average correlation 0.63), with similar performance for Hispanics and Asians, somewhat lower performance for whites, and notably lower performance for blacks. GO (average correlation 0.53) was near the performance of the BSG and notably better than CSG for blacks, but performed less well than the other two algorithms for all other groups, performing especially poorly for Hispanics and Asians. Overall, BSG was 74 percent more efficient than geocoding alone in estimating individual race/ethnicity and CSG was 41 percent more efficient than geocoding alone in predicting individual race/ethnicity. This means that 1,000 observations from BSG provide as much information as 1,740 observations using geocoding alone. For Hispanics and Asians, BSG has 2.6 and 3.9 times the efficiency of geocoding alone, respectively.

BSG performed better than each of the alternatives by both performance metrics and increases efficiency by 56–74 percent relative to geocoding alone. In contrast, the CSG improves upon direct use of geocoded data by only one of these metrics, highlighting the importance of *how* surname and geocoded information is combined.

Discussion

We have described a method for estimating race/ethnicity using administrative data. This approach, which applies Bayes's Theorem to a four-category geocoding and surname analysis, appears to be a particularly useful means of integrating these sources of information and substantially outperforms a classification-based means of combining this information (CSG). The advantage of BSG over CSG probably stems from two factors: (1) better identification of blacks in areas of low residential segregation and (2) greater precision through the direct use of probabilities.

In addition to its ability to estimate race/ethnicity, the BSG approach has substantial potential for use in routine assessment and monitoring of health disparities in a population. It can also be used when estimated race/ethnicity is to be a predictor in multivariate regression or other models; thus its usefulness is not limited to estimation of disparities or to health applications.

One limitation, which applies to all methods of inferring race/ethnicity, is that while BSG supports modeling at the individual level, it is not accurate enough to support individual-level interventions and

requires large sample sizes for good precision, because there is some inherent loss of information compared with self-reported race/ethnicity for a sample of the same size. Secondly, although results were evaluated on a large, racially and ethnically diverse national sample, results may differ somewhat for those not insured by this health plan or those who do not self-report race-ethnicity.

An additional limitation is that the direct use of predicted probabilities is somewhat more complex than the use of 1/0 categorical indicators of race/ethnicity and may be unfamiliar to some analysts. Traditionally, analysts have either used a single categorical variable with each level representing a particular racial/ethnic group, or a series of "dummies," that is—separate variables (one for each race/ethnicity) that have a value of "0" if the person is not, for example, Asian, or "1" if the person is Asian. The posterior probabilities from the BSG and GO are continuous variables with values from 0 to 1 that are used somewhat differently. Nonetheless, this approach is still relatively straightforward, and one can interpret the coefficients as if they were from racial/ethnic dummy variables. The Appendix S1 provides examples of how these probabilities can be used within *SAS*.

Our new method of estimating race/ethnicity substantially outperforms other widely used indirect methods and provides health plans and others a timely means to infer race/ethnicity among plan members. Although self-reported race/ethnicity represents a gold standard in many situations, indirect methods offer a powerful and immediate alternative for estimating health experiences by racial/ethnic status using only administrative data. In combination with geographic information systems (GIS) tools, these methods can be of great use to health plans, researchers, and others (National Health Plan Collaborative 2006).

Future work can directly examine the accuracy of BSG in estimating health disparities, as well as seek further improvements in the accuracy of BSG estimates of race/ethnicity. One way to do the latter might be to develop regional sensitivity and specificity parameters. Such data would also provide insight into the extent to which BSG performance varies by plan or region. One could model racial-ethnic selection into health insurance within Block Group conditional on surname results, further improving BSG performance (because our results imply there are lower rates of health coverage for blacks and Hispanics than for Asians and whites/others even within the same Block Groups).

Finally, when applying BSG to a specific population, such as a commercially insured population, one could use Census racial/ethnic data within block groups that were restricted to ages that better matched the target population. To the extent that age differed by race/ethnicity, this would further reduce BSG bias and improve its performance. Future work should follow along these paths to refine an already promising and useful approach to inferring race/ethnicity from names and addresses alone.

Acknowledgments

This study was supported, in part, by contract 282-00-0005, Task Order 13 from DHHS: Agency for Healthcare Research and Quality. Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056). The authors would like to thank Kate Sommers-Dawes and Scott Stephenson for assistance with the preparation of the manuscript.

Disclaimers: The contents of the publication are solely the responsibility of the authors and do not necessarily reflect the official views of the CDC.

Disclosure: None

Supplementary Material

The following material is available for this article online:

Appendix S1: Implementation of Bayesian Surname and Geocoding Combination (BSG).

Appendix S2: Author Matrix.

Table S1: Probabilities of Joint Surname Test Results by True Race/Ethnicity.

Table S2: Posterior Probabilities of GroupMembership by Surname List Results.

Table S3: BSG Posterior Probabilities of Race/Ethnicity for Hypothetical Block Groups A, B, C, D, and E for Males (N5963,319).

Table S4: Example of BSG, GO, and CSG estimates of Racial/Ethnicity of PlanMembership in a Hypothetical Block Group (67 Percent White/Other, 11 Percent Each Black, Hispanic, and Asian Overall), Based on 10 Male# Plan Members (2 on Asian Surname List, 3 on Spanish Surname List, 5 Unlisted).

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2008.00854.x (this link will take you to the article abstract).

Click here to view.(68K, doc)

Click here to view.(247K, pdf)

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

Abrahamse AF, Morrison PA, Bolton NM. Surname Analysis for Estimating Local Concentration of Hispanics and Asians. Population Research and Policy Review. 1994;13:383–98. [Google Scholar]

Boston Public Health Commission. Data Collection Regulation. Boston: Boston Public Health Commission; 2006. [Google Scholar]

Brier GW. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review. 1950;78(1):1–3. [Google Scholar]

California State Senate. Senate Bill Analysis of SB 853. Sacramento, CA: California State Senate; 2007. [Google Scholar]

Falkenstein MR. The Asian and Pacific Islander Surname List: As Developed from Census 2000. 2002. Paper read at Joint Statistical Meetings.

Fiscella K, Fremont AM. Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity. Health Services Research. 2006;41(4, pt 1):1482–500. [PMC free article] [PubMed] [Google Scholar]

Fremont AM, Lurie N. The Role of Race and Ethnic Data Collection in Eliminating Health Disparities. Washington, DC: National Academies Press; 2004. [Google Scholar]

Fremont AM, Pantoja P, Elliott MN, Morrison PA, Abrahamse AF, Lurie N. Use of Indirect Measures of Race/Ethnicity to Examine Disparities in Managed Care. 2005. AcademyHealth Annual Research Conference. Chicago, IL.

Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: The National Academies; 2002. [PMC free article] [PubMed] [Google Scholar]

Kestenbaum BB, Ferguson R, Elo I, Turra C. Hispanic Identification. 2000. Paper read at 2000 Southern Demographic Association Meetings.

Lauderdale D, Kestenbaum BB. Asian American Ethnic Identification by Surname. Population and Development Review. 2000;19(3):283–300. [Google Scholar]

McCaffrey D, Elliott MN. Power of Tests for a Dichotomous Independent Variable Measured with Error. Health Services Research. 2007 DOI 10.1111/j.1475-6773.2007.00810.x. [PMC free article] [PubMed] [Google Scholar]

Morrison PA, Kestenbaum B, Lauderdale DS, Abrahamse AF, El-Badry S. Developing an Arab-American Surname List: Potential Demographic and Health Research Applications. 2003. Paper read at 2003 Southern Demographic Association Meetings.

National Health Plan Collaborative. Phase 1 Summary Report: Reducing Racial and Ethnic Disparities and Improving Quality of Health Care. Hamilton, NJ: National Health Plan Collaborative; 2006. [Google Scholar]

National Research Council. Eliminating Health Disparities: Measurement and Data Needs. Washington, DC: National Academies Press; 2004. [Google Scholar]

Perkins RC. Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results. 1990. U.S. Census Bureau, Population Division.

Yates JF. External Correspondence: Decompositions of the Mean Probability Score. Organizational Behavior and Human Performance. 1982;30:132–56. [Google Scholar]

Assessing Fair Lending Risks Using Race/Ethnicity Proxies

Yan Zhang

Enterprise Risk Analysis Division

Office of the Comptroller of the Currency

Washington, DC 20219

yan.zhang@occ.treas.gov

Abstract

Fair lending analysis of non-mortgage credit products often involves proxying for race/ethnicity since such information is not required to be reported. Using mortgage data, this paper evaluates a series of proxy approaches (geo, surname, geo-surname, and BISG) as compared with the race/ethnicity reported under HMDA. The BISG proxy predicts the reported race/ethnicity the best as judged by prediction bias, correlation coefficient, and discriminatory power. In assessing fair lending risks where classification of race/ethnicity is called for, we propose the BISG maximum classification, which produces a more accurate estimation of mortgage pricing disparities than the current practices. The above conclusions withhold various robustness tests. Additional analysis is performed to assess the proxies on non-mortgage credits by leveraging consumer credit bureau data.

Keywords: fair lending risk, race/ethnicity, proxy, BISG, Bayesian, measurement error, misclassification.
JEL classifications: C11, C81, D18, J15.

I. Introduction

The Equal Credit Opportunity Act (ECOA) prohibits a creditor from discriminating against any borrower on the basis of race, color, religion, national origin, sex, marital status, or age. Under ECOA, regulatory agencies assess fair lending risks of lending institutions by comparing lending outcomes based on the above-mentioned prohibited basis factors. Failure to comply with ECOA can subject a financial institution to civil liability for actual and punitive damages in individual or class actions.[1]

Consumer lending products can be categorized into two groups: mortgage and non-mortgage products. Common non-mortgage products are credit card, auto loan, student loan, consumer loan, and small business loan. Historically, fair lending analysis and research have been more focused on mortgage than non-mortgage loans. One important reason for this is the availability of accurate data on prohibited basis factors. The Home Mortgage Disclosure Act (HMDA) authorizes lenders to collect information on the race, ethnicity, and gender of mortgage applicants and co-applicants. However, lenders generally are not permitted to collect such information for non-mortgage products.

Nonetheless, the fair lending evaluation of non-mortgage credit products is indispensable in ensuring that a lending institution is fully compliant with ECOA. Since its inception in July 2011, the Consumer Financial Protection Bureau (CFPB) has made multiple Department of Justice (DOJ) referrals on non-mortgage lending, especially indirect auto lending. In March 2013, the CFPB issued Bulletin 2013-02 regarding fair lending risks in indirect auto finance. As a result of the CFPB referral, Ally Bank was ordered to pay $80 million in damages to harmed applicants in December 2013. The proportion of non-mortgage referrals has also been increasing overall. Among the 18 referrals made by regulatory agencies to the DOJ for violation of ECOA in 2014, 15 involve discrimination in non-mortgage lending.[2]

The CFPB and DOJ used race/ethnicity proxies to estimate disparities in dealer markups on the basis of race and national origin at Ally Bank. Since mid-1990, proxy methods have been used in fair lending analysis when self-reported prohibited basis factors are not available (Baines and Courchane, 2014), leveraging findings from other fields, mainly epidemiology. For race/ethnicity, which this paper focuses on, the common approaches are surname and geocoding (or simply geo) methods, which use the Census surname list or geographic composition to impute race/ethnicity. Fiscella and Fremont (2006) provide a comprehensive review of literature that uses geocoding and surname approaches to assess disparities in health care. In addition, hybrid approaches have been suggested that attempt to create a more refined measure using both pieces of information. The most recent and advanced hybrid approach is the Bayesian Improved Surname Geocoding method (BISG) by Elliott et al. (2009), which is the proxy

[1] The Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 (Dodd-Frank Act) granted rule-making authority under ECOA to the Consumer Financial Protection Bureau (CFPB).
[2] The Attorney General's 2014 Annual Report to Congress Pursuant to the Equal Credit Opportunity Act Amendments of 1976, April 2015.

method adapted by CFPB and DOJ in the Ally Bank case. The BISG method aims to improve upon the information in the Census surname list by incorporating the race/ethnicity composition of the applicant's neighborhood of residence. Elliott et al. (2009) use health plan enrollment data with reported race/ethnicity to evaluate how well the BISG method predicts race/ethnicity and find that the BISG methodology is more accurate than the geo-only or surname-only approach. Simpler approaches to integrate multiple sources of information have also been proposed and evaluated. Coldman et al. (1988) compare four rules to combine first name, middle name, and surname to identify Chinese vs. non-Chinese. They find that the linear rule, which averages the three individual probabilities of name components, achieves the best sensitivity and specificity.

Assessing fair lending risks using imputed race/ethnicity is a timely topic, but relevant research is rather recent and still limited. Attempting to evaluate redlining risk of credit cards, Cohen-Cole (2011) links the location-based information on race to an individual borrower's access to credit cards. The CFPB (2014) evaluates the BISG methodology using mortgage data and concludes that the BISG proxy is more accurate than the geo-only or surname-only proxy in predicting an individual applicant's reported race/ethnicity. The CFPB study evaluates the accuracy of BISG over other proxies as measured by correlation coefficient and Gini index, but not its impact on fair lending risk assessment. Baines and Courchane (2014) claim that the BISG methodology produces biased race/ethnicity proxies, which result in inflated differences in assessing lending disparities. Using a mortgage dataset, they show that the classification errors of BISG are correlated with creditworthiness factors and conduct regression analysis to estimate the raw mortgage pricing disparities using the reported race/ethnicity vs. the BISG proxy. However, despite acknowledging the importance of considering creditworthiness factors, Baines and Courchane's regression analysis does not control for legitimate factors that lenders consider in mortgage pricing decisions, which weakens their conclusion.[3]

The BISG proxy produces a continuous value between 0 and 1 as the probability of the borrower belonging to a specific race/ethnicity, with the probabilities of all race/ethnicity groups summing up to 1. Since fair lending assessment aims to estimate the lending differences between a particular prohibited basis group (PBG) and the control group (CG), borrowers need to be assigned to one of the race/ethnicity groups exclusively and completely, which requires the continuous BISG probability to be dichotomized into a binary race/ethnicity indicator. Practically, there have been two approaches adopted by fair lending analysis so far: one is to dichotomize the BISG probability using a fixed threshold (Baines and Courchane, 2014), while the other is to use the continuous probability as it is (CFPB, 2014). We term the former approach the BISG fixed and the latter the BISG continuous. Since CFPB (2014) does not

[3] Baines and Courchane (2014) do evaluate the adjusted disparities using auto lending data, but as the data do not contain reported race/ethnicity, the potential bias caused by using the BISG proxy cannot be directly validated.

describe whether they conduct a regression analysis or how they use the BISG continuous in the regression setting, we take the liberty to assume that regression analysis using the BISG continuous estimates the disparities between all PBGs against the CG simultaneously in one equation. Conducting separate regressions by PBG is not feasible because the BISG continuous does not assign a borrower to a definitive race/ethnicity group. The one regression approach adopted by the BISG continuous restricts the other explanatory variables in the regression to be of the same effect for all PBGs, which is an assumption often too strong to be valid because loan and borrower characteristics tend to be correlated with race/ethnicity. Moreover, the BISG continuous is subject to estimation bias: The population with a specific credit product could be different from the U.S. population on which the reference surname and geo databases are based; and the calculation of BISG probabilities assumes conditional independence, which is often not satisfied in reality. Since the BISG fixed uses the BISG continuous as an input, the bias of the BISG continuous can transit into the misclassifications of the BISG fixed. The fixed classification approach itself might cause additional misclassifications. The BISG fixed requires a threshold to classify the continuous probability into a binary indicator. False positives decrease with the threshold, but false negatives increase with the threshold. If the threshold is too high, a borrower might not be assigned to any of the race/ethnicity groups; if the threshold is too low, a borrower might end up in more than one race/ethnicity group.

In the field of machine learning, the "maximum a posteriori" (MAP) decision rule is recommended for the BISG type of naive Bayesian probability model, as naive Bayesian probability often reserves the rank ordering of the probabilities among classes despite having estimation errors. Moreover, not dependent on a fixed threshold, the MAP rule ensures that each probability is assigned to one class exclusively and completely. Leveraging the expertise from machine learning, we propose the BISG maximum classification (BISG max), which assigns a borrower to the race/ethnicity with the highest BISG probability and compare it with the BISG continuous and BISG fixed threshold classification.

This paper contributes to the fair lending literature by providing a comprehensive review of the race/ethnicity proxy methods and a rigorous analysis of their suitability for and limitations in fair lending risk assessment. Using mortgage data, this paper compares a series of proxies with the race/ethnicity reported under HMDA. In addition to the geo, surname, and BISG proxies, it also considers the linear rule (Coldman et al., 1988), which combines the geo and surname information by taking their average.[4] Our analysis shows that the BISG produces the most accurate estimates of race/ethnicity probabilities among the four proxies, as judged by estimation bias, correlation coefficient, and discriminatory power of the

[4] Elliott et al. (2008) have already shown that the Bayesian Surname Geocoding (BSG) proxy, which is the predecessor of BISG, performs better than the Categorical Surname and Geocoding (CSG) proxy (Fremont et al., 2005). The CSG proxy uses either the surname or the geographic information to categorize race/ethnicity in a sequential order of Hispanic, Asian, Black, and non-Hispanic White.

reported race/ethnicity. By merging HMDA with DataQuick, we obtain important factors typically considered in mortgage pricing decisions[5] to enable a more comprehensive regression analysis. Our analysis shows that the BISG max greatly reduces the estimation bias in disparity coefficient as compared with the BISG continuous and BISG fixed, which holds true under a series of robustness analyses.

Furthermore, as a remedy to the lack of non-mortgage data, this paper links mortgages with non-mortgages using the credit bureau data (CBD) of the Office of the Comptroller of the Currency (OCC) to shed light on non-mortgages. Among the CBD borrowers who have at least one form of major credits (we consider mortgage, credit card, auto loan, and student loan), about 38% of the CBD borrowers have at least one mortgage, and almost all of them have one or more of the other three credit products. Therefore, despite being a population conditional on having a mortgage, the linked non-mortgage population covers a great portion of the non-mortgage universe. Though regression analysis is not feasible due to data limitations, the non-mortgage analysis replicates the univariate results of the mortgage analysis: The BISG is a better predictor of reported race/ethnicity than the geo, surname, and geo-surname proxies, and the BISG max has a better coverage than the BISG fixed.

This paper also adds to the research on dichotomization of mismeasured predictors. Using race/ethnicity proxies in fair lending regression analysis presents a complicated and yet intriguing case for such study. First, the continuous race/ethnicity probabilities are expected to have measurement errors (and moreover) with bias. Second, fair lending regression analysis often accounts for various legitimate factors, with which race/ethnicity could be correlated. Third, there are multiple race/ethnicity probabilities associated with one borrower. Fourth, separate regression is preferred to one general regression to allow the effect of other covariates to vary by PBG. This paper provides an empirical study that compares the effects of continuous covariate mismeasurement vs. dichotomized covariate misclassification involving the above-mentioned challenges. To the best of our knowledge, there has not been such a discussion in the existing relevant literature. McCaffrey and Elliott (2008) examine the efficiency of using predicted probability for a binary independent variable in predicting continuous outcome and recommend direct substitution rather than classification. Allowing for another predictor variable that is potentially correlated with the predictor of interest, Gustafson and Le (2002) find that dichotomization can actually reduce the estimation bias due to mismeasured covariate in some instances. However, both studies assume that there is only one dichotomous covariate and it is measured without bias; these conditions are not satisfied in the fair lending risk assessment.

Additionally, this paper contributes to the Bayesian literature by presenting an empirical application of Bayes' rule. Comparing the BISG proxy that uses Bayes' rule with the linear approach

[5] Typical lending outcomes include underwriting and pricing. We focus on pricing disparity analysis because DataQuick data only report originated mortgage loans.

illustrates the benefit of using the more sophisticated BISG approach. This paper introduces the MAP rule from the field of machine learning into fair lending analysis. It confirms empirically that the max classification of naive Bayesian probability also leads to optimal results in assessing fair lending risks.

The rest of the paper is organized as follows. Section II discusses the race/ethnicity proxy methods: geo, surname, geo-surname, and BISG. Section III describes the mortgage data used. Section IV compares the performance of the continuous race/ethnicity proxies in predicting HMDA reported race/ethnicity. Section V lays out the classification methods. Section VI evaluates the mortgage pricing disparities using the proposed BISG max classification and existing approaches. Section VII assesses the proxies on non-mortgage data. Section VIII concludes the paper.

II. The Proxy Methodologies for Race/Ethnicity

The main methods to proxy for race/ethnicity can be summarized into two groups and four types: the single-sourced geocoding and surname approaches, and the hybrid geo-surname and BISG approaches.

II.A. Geocoding and Surname Methods

The geocoding method infers an individual's race/ethnicity based on where he/she lives. The Census data provide socioeconomic status aggregated at various geographic area levels, among which is race/ethnicity. By linking the individual to the Census database, the prevalence of the race/ethnicity shown by the Census data is used as the probability of the individual belonging to the corresponding race/ethnicity. The geocoding approach can be mathematically expressed as the following

$$Geocoding: p(r|g) = \frac{N_{rg}}{N_g}, \quad (1)$$

in which N_g is the number of people in the geographic area g where the individual lives, and N_{rg} is the number of people belonging to race/ethnicity r in the same area g. Geocoding is found to be effective in more segregated areas, where one race/ethnicity has a dominantly high concentration, and less predictive in integrated areas, where multiple races/ethnicities coexist without a dominant one. Among the various race/ethnicity categories, Blacks tend to live in more segregated areas; therefore, geocoding is deemed to be more powerful in identifying Blacks (Fiscella and Fremont, 2006).

The surname approach generates a probability of race/ethnicity using the surname of an individual. By matching the individual's surname to an existing surname database with a corresponding percentage for each race/ethnicity, the race/ethnicity percentage is adopted as the race/ethnicity prediction of the individual. The calculation for the surname approach is

$$Surname: p(r|s) = \frac{N_{rs}}{N_s}, \quad (2)$$

where N_s is the number of people nationwide with surname *s*, and N_{rs} is the number of people of race/ethnicity *r* with surname *s*. In order to produce a high confidence in a race/ethnicity proxy, the surname needs to be highly associated with a certain race/ethnicity. Empirically, surnames of Hispanics and Asians are relatively easy to be identified due to their uniqueness, leading to higher imputation accuracy for them by this approach (Fiscella and Fremont, 2006).

II.B. The Geo-Surname and BISG Methods

As the prediction accuracy of geocoding and surname methods is limited to certain races/ethnicities, hybrid approaches that can integrate both the location and surname of an individual are proposed to generate more accurate race/ethnicity predictions across all race/ethnicity groups. Coldman et al. (1988) compare four rules (multiplicative, linear, maximum, last name) in integrating first name, middle name, and last name to identify the Chinese ethnic group and find that the linear rule performs best. Similarly, we test the linear rule in combining both the location and surname of a borrower, which, denoted as the geo-surname proxy, is constructed as

$$Geo - surname = \frac{p(r|g)+p(r|s)}{2}. \quad (3)$$

Elliott et al. (2008) develop the Bayesian Surname Geocoding (BSG) method, which uses a Bayesian approach to combine surname and geocoded information. The Bayesian method consists of two parts. The first is the creation of a "prior" probability of an individual belonging to a race/ethnicity using the surname information. The second generates the "posterior" probability, which involves updating the "prior" probability using the information on the demographic characteristics of the person's place of residence. Elliott et al. (2009) further refine the BSG method by proposing the Bayesian Improved Surname Geocoding (BISG) method. The BISG method improves over the previous BSG method by using a more recent surname list to create the prior probability and using more categories of race and ethnicity,[6] but the information integration technique remains the Bayesian approach.

Using the same notations as those for surname or geocoding proxy, let *r* denote race, *s* denote surname, and *g* denote geography or location. The prior probability is the person's race based on his/her surname and can be written as $p(r|s)$, where *r* takes on values 1 through 6 for each of the six mutually exclusive races: Hispanic, non-Hispanic White, non-Hispanic Black or African American, non-Hispanic Asian/Pacific Islander (API), non-Hispanic American Indian and Alaska Native (AIAN), and non-

[6] The surname list used by the BSG is derived from the 1990 Census, while the BISG method uses the list derived from Census 2000. The BSG method uses a four-category race/ethnicity definition, compared with the six categories used by BISG.

Hispanic multiracial.[7] This probability is then updated using the race/ethnicity composition of the place of residence. The posterior probability $p(r|g,s)$ is the BISG proxy. Using Bayes' rule and the chain rule, the BISG proxy can be written as

$$BISG: p(r|g,s) = \frac{p(r,g,s)}{p(g,s)} = \frac{p(s)p(r|s)\,p(g|r,s)}{\sum_r p(s)p(r|s)p(g|r,s)} = \frac{p(r|s)\,p(g|r,s)}{\sum_r p(r|s)p(g|r,s)}, \quad (4)$$

where $p(g|r,s)$ is the probability of a borrower residing in a certain location given race *r* and surname *s*. However, there is no nationwide public database for $p(g|r,s)$, so Elliott et al. (2008) introduce the conditional independence assumption to enable the calculation of Eqn. (4). The assumption is that the probability of an individual residing in a certain location given a person's race does not vary by surname; i.e., $p(g|r,s) = p(g|r)$. Different from the geocoding probability $p(r|g)$, the term $p(g|r)$ is calculated as the proportion of the entire U.S. population with race *r* in location *g*

$$p(g|r) = \frac{N_{rg}}{N_r}, \quad (5)$$

where $N_{r\cdot}$ is the U.S. population with race *r*, and N_{rg} is the number of residents with race *r* in location *g*. Assuming conditional independence and plugging Eqn. (2) and (5) into Eqn. (4), the BISG estimator can be simplified as:

$$BISG\ simplified: p(r|g,s) = \frac{p(r|s)\,p(g|r)}{\sum_r p(r|s)p(g|r)} = \frac{\frac{N_{rs}}{N_s}\frac{N_{rg}}{N_r}}{\sum_r \frac{N_{rs}}{N_s}\frac{N_{rg}}{N_r}}. \quad (6)$$

The BISG algorithm requires a geo-match to run, which is typically satisfied. When a surname does not match to the Census surname list, the BISG method first imputes the probability of race/ethnicity given the surname, $p(r|s)$, using the national average of race/ethnicity. It then combines that with the geographic location information to calculate the final prediction using Bayesian probability theory. Theoretically, the resulting BISG prediction is the same as the geo-only prediction.[8]

II.C. The Geocoding and Surname Databases

The BISG method uses the Census 2010 Summary File 1 (SF1) for calculating $p(g|r)$ and Census 2000 surname list to extract $p(r|s)$, as these two datasets are the most recent and comprehensive data sources for tabulating surname and geographic area with race/ethnicity. They are the same data sources that will be used by other proxy methods discussed here.

[7] For simplicity, we omit "non-Hispanic" when referring to race throughout the rest of the paper; for example, White means non-Hispanic White.

[8] Let *N* denote the total population of the U.S., so $\frac{N_r}{N}$ is the national average of race *r*; inserting $\frac{N_{rs}}{N_s} = \frac{N_r}{N}$ into Eqn. (6), the BISG prediction for race *r* given *g* for surname non-matches is $BISG\ Surname\ nonmatch: p(r|g,s) = \frac{\frac{N_r}{N}\frac{N_{rg}}{N_r}}{\sum_r \frac{N_r}{N}\frac{N_{rg}}{N_r}} = \frac{N_{rg}}{\sum_r N_{rg}} = \frac{N_{rg}}{N_g}$, which reduces to the geo-only method.

The Census 2010 SF1 provides population counts by race/ethnicity across geographic regions at six levels (from largest to smallest): nation, state, county, tract, block group, and block. Imputation of race/ethnicity at a more granular geographic level enables capturing local level segregation, but on the other side, the smaller population for race/ethnicity imputation could impact the robustness of the race/ethnicity distribution. Based on the 2010 Census, there are 73,057 tracts, 217,740 block groups, and 11,078,297 blocks, with average populations of roughly 4,200, 1,200, and 28 respectively.[9] The Census tract and block group are commonly used in geocoding race/ethnicity, as they contain a reasonably sufficient number of residents and yet are not too large.[10] We present the results using Census SF1 data at the Census tract level. We also evaluate the proxies using block group SF1 data (details are provided in Appendix 1 of the e-companion[11]), and the results are almost identical to the tract level results.

The Census surname database is based on Census 2000, which was released by the Census Bureau in 2007. The Census surname database contains 151,671 surnames listed by 100 or more individuals and represents about 89.8% of all individuals enumerated in Census 2000. For each surname, the surname database provides the percentage of individuals who belong to one of six mutually exclusive and exhaustive race/ethnicity categories: Hispanic, White, Black, API, AIAN, and multiracial.

In order to assign probabilities using the above-mentioned surname and geography databases, the dataset needs to be prepared so that it can be matched to the Census databases. Elliott et al. (2009) develop SAS programs to clean and standardize[12] the surnames as a part of their BISG algorithm, which we leverage in our analysis. If the dataset does not contain the matching variables (such as Census tract or block group), geocoding software[13] can be used to obtain such information using the exact street address as the input.

III. Mortgage Data

We first use mortgage data to evaluate race/ethnicity proxies because race/ethnicity is self-reportable under HMDA.[14] The data consist of mortgages originated for the purpose of home purchase

[9] Source: https://www.census.gov/geo/maps-data/data/tallies/tractblock.html.
[10] Elliott et al. (2008, 2009) use Census SF1 at the block group level and go to the next higher geographic level (tract) if block group geocoding is not feasible; similarly, CFPB (2014) adopts a geocoding hierarchy composed of block group, tract, and five-digit zip code; and Baines and Courchane (2014) use Census demographics at the tract level.
[11] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.
[12] For example, removing special characters and titles, and parsing hyphenated or compound names.
[13] For example, ArcGIS (http://www.esri.com/).
[14] We acknowledge that reported race/ethnicity is potentially subject to reporting errors. But as this paper focuses on the accuracy of race/ethnicity proxies, we assume the reported race/ethnicity is true to the real race/ethnicity and use it as the benchmark for proxy comparison.

from 2004 through 2007 in 10 major Metropolitan Statistical Areas (MSAs) of the U.S.[15] The 10 MSAs are selected so that the data cover major U.S. MSAs with various race/ethnicity compositions. We focus on originations of this time period because HMDA did not adopt the current race/ethnicity definition until 2004. Moreover, as we leverage the DataQuick private label securitized (PLS) loan database, ABS/MBS, to extract more detailed loan level information, we would like to select a time period for which PLS loans are more representative of U.S. mortgage loans. As Figure 1 shows, the private label securitization of residential mortgage loans went through unprecedented growth during 2004–2007, such that its loan volume became comparable to or exceeded that of the government-sponsored enterprises (GSEs). It began to decrease starting in 2007Q3 and became negligible after 2007.

To obtain surnames and locations needed to derive the proxies, we leverage DataQuick Property data. The Property database reports data compiled daily from local county recorder transactions on residential and commercial properties nationwide, including the full names and addresses of those listed on the title. We use the surnames of sole applicants or primary applicants if the mortgage is owned jointly. The addresses are geocoded to obtain the 2010 Census tracts. The Origination table of the DataQuick ABS/MBS data lists important information at loan origination, including FICO, loan-to-value (LTV) ratio, combined LTV (CLTV) ratio, interest rate in the form of note rate, origination date, collateral type, loan product, etc. The DataQuick ABS/MBS data contain nearly 23 million loan originations, consisting of about 95% of the non-agency securitized securities.[16]

Through a meticulously crafted matching scheme, the analysis dataset is created by first merging HMDA with the Property data of DataQuick. Specifically, a match is declared between HMDA and DataQuick if they match exactly on loan amount, Census tract, loan type, and their closing dates are no more than 30 days apart. If multiple matches are found after imposing the above matching criteria, lender name is used to further narrow down. To be conservative, all the multiple matches are dropped if comparing lender names cannot produce a unique match. As some banks report to OCC with HMDA-plus data that append borrowers' names to regular HMDA, we leverage such HMDA-plus data to examine the quality of HMDA-DataQuick (HMDA-DQ) matching. Based on the HMDA-plus data of a particular financial institution over a specific time period, we find a very low error rate of 0.6% for the name matching. Fully acknowledging its limited coverage, this validation analysis lends us comfort to the quality of the HMDA-DQ matching algorithm.[17] The HMDA-DQ matched sample is then further

[15] The 10 MSAs are 12060 (Atlanta-Sandy Springs-Marietta, GA), 14460 (Boston-Cambridge-Quincy, MA-NH), 16980 (Chicago-Naperville-Joliet, IL-IN-WI), 19820 (Detroit-Warren-Livonia, MI), 31080 (Los Angeles-Long Beach-Anaheim, CA), 33100 (Miami-Fort Lauderdale-Miami Beach, FL), 35620 (New York-Northern New Jersey-Long Island, NY-NJ-PA), 38300 (Pittsburgh, PA), 41740 (San Diego-Carlsbad-San Marcos, CA), and 41860 (San Francisco-Oakland-Fremont, CA).
[16] Source: http://www.mbsdata.com/datasets_new.html.
[17] For complete details of the data merge, please refer to appendix A of Mayock and Spritzer (2015).

enriched with detailed origination information by linking to the DataQuick ABS/MBS data through DataQuick internal property ID and origination date.

The final HMDA-DQ data contain 630,807 observations with non-missing surnames and Census tracts as well as origination information. All the 630,807 loans can be linked to the Census SF1 via Census tracts. We then ran the surname cleanup program and matched the cleaned surnames to the Census 2000 surname database. The overall surname matching rate of the HMDA-DQ data is 86.3%, which is quite close to the 89.8% matching rate of the Census surname database. In the range of 83-90%, the surname matching rate is fairly stable by the race/ethnicity category, MSA, or loan origination year.

Panel A of Table 1 contains the self-reported race/ethnicity composition of the HMDA-DQ data, as compared with the U.S., the 10-MSA, and the 10-MSA adult (with age equal to or older than 18) populations. We report four out of the six race/ethnicity categories (Hispanic, Black, API, and White) since AIAN and multiracial are very limited in the U.S. The 10-MSA population contains a high proportion of Hispanic, Black, and API, and a lower proportion of White, as compared with the entire U.S. population. Imposing the age restriction alleviates the difference to some extent, as shown by the race/ethnicity composition based on the 10-MSA adult population. The HMDA-DQ dataset is composed of 33.0% Hispanic, 11.5% Black, 9.9% API, and 45.1% White, providing sufficient numbers of racial/ethnic minorities to achieve a robust evaluation of proxy approaches. On the other hand, given the noticeable differences in the race/ethnicity composition between the Census and the HMDA-DQ data, the latter is unlikely to be a random sample of the former and therefore the race/ethnicity proxies constructed for the analysis data using Census information are subject to bias.

Panels B and C of Table 1 show that the analysis dataset covers various levels of race/ethnicity composition across MSAs and over time. As much as 52.1% of the population in the Miami MSA (33100) is Hispanic; the Atlanta MSA (12060) shows a high concentration of Black (45.0%); the San Francisco MSA (41860) has 24.2% API; and the Pittsburgh MSA (38300) is predominantly (87.5%) occupied by White. We observe a noticeable increase in purchases by White and API, and a similar decrease by Hispanic and Black in 2007, which coincides with the beginning of the subprime crisis.

IV. Comparing BISG with Other Proxies

In this section we compare the performance of BISG with the geo, surname, and geo-surname methods in predicting HMDA reported race/ethnicity. Since the direct output of all four proxies is a continuous race/ethnicity probability for a borrower, the comparison measures here are prediction bias, correlation coefficient, and discriminatory power, which are common statistics used to evaluate a continuous probability prediction of a binary outcome.

IV.A. Prediction Bias

Panel A of Table 2 lists the prediction bias of the BISG, geocoding, surname, and geo-surname for each race/ethnicity, which is calculated as the difference between the mean predicted probability of the proxy and the average of reported race/ethnicity indicator. Often, the prediction bias deviates from zero, which confirms that the proxies are biased, as discussed in Section III. In particular, the BISG shows the smallest bias and the geocoding the largest bias for Hispanic; the BISG predicts Black with the best accuracy and the surname the worst; for API and White, the geocoding is the most accurate and the surname is the worst. The prediction bias of the geo-surname falls between (is the average, to be exact) that of the geo's and the surname's because the geo-surname is an arithmetic average of the geo and surname.

IV.B. Correlation Coefficient

The correlation coefficient is used to measure how related two variables are. The correlation coefficient takes a continuous value from -1 to 1, and the closer its absolute value is to 1, the more related the two variables are. The sign of the correlation coefficient indicates the direction of the relationship, with positive indicating the two variables move in the same direction and negative the opposite. The Pearson correlation coefficient detects a linear relationship between two variables, and a rank correlation such as Spearman evaluates an ordinal relationship.

Panel B of Table 2 displays the Pearson correlation coefficient between each proxy and the self-reported race/ethnicity.[18] There are several important findings. First, the BISG and geo-surname proxies are more highly correlated with the self-reported race/ethnicity than the geo-only and surname-only proxies across all four race/ethnicity categories. In addition, most correlation coefficients of the hybrid approaches are high, with absolute values above 70%, suggesting that they are reasonably well correlated with the actual values. This finding attests the benefit of considering both location and surname of a borrower. Second, between the two hybrid approaches, the correlation coefficients of the BISG are higher than those of the geo-surname, showing the advantage of using the more sophisticated Bayesian approach over the simple averaging approach. Third, with the exception of Black, the surname-only proxy has a higher correlation coefficient than the geo-only proxy. This is consistent with the prior assessment of the geo-only proxy that its effectiveness is limited to areas with a dominant race/ethnicity, usually Black.

IV.C. Discriminatory Power

[18] Spearman correlation analysis leads to the same conclusion and is not provided here for brevity.

The Receiver Operating Curve (ROC) illustrates the discriminatory power of a probability prediction as its classification threshold varies from 0 through 1. It is created by plotting the true positive rate (sensitivity) on the y-axis against the false positive rate (1-specificity) on the x-axis for every possible value of the probability that the threshold may take. Each point on the ROC represents a trade-off between sensitivity and specificity. A perfect prediction will have a 100% true positive rate without any false positives; therefore, an ideal ROC would run vertically from the origin (0,0) upwards to point (0,1), and then run horizontally from there to point (1,1). The worst ROC is the 45 degree diagonal line, indicating that the corresponding prediction does not have any discriminatory power as it assigns equal chances of true positives and false positives. The Area under the Curve (AUC) is a quantitative measure of the discriminatory power of a probability prediction. Note that another commonly used measure of discriminatory power, the Gini coefficient, is just a monotonic transformation of AUC, which can be calculated as (2×AUC-1). The higher the AUC or Gini coefficient is, the better a predictor can classify a binary outcome.

Figure 2 presents the ROC and AUC for each race/ethnicity. Consistent with the pattern found in the correlation analysis, the BISG has the highest AUC (all above 92%) among the proxies, followed by the geo-surname, the surname-only, and finally the geo-only, whose performance is the weakest among the four. Statistical tests show that under each race/ethnicity, the AUC of the BISG is higher than any of the other three methods at the 1% statistical significance level.

To summarize, the BISG produces the most accurate prediction of the reported race/ethnicity among the four proxies as measured by prediction bias, correlation coefficient, and discriminatory power.

V. Classification

As fair lending analysis compares lending differences between one particular PBG and the CG (which is usually White), it calls for a classification that dichotomizes the continuous probability of race/ethnicity into a binary indicator of race/ethnicity. We compare the classification methods using the BISG probability as the underlying continuous variable, since it has demonstrated better performance than the other three proxies in predicting race/ethnicity under a series of performance measures in Section IV.

A common approach to transforming a continuous probability into a binary indicator of race/ethnicity is to use a fixed threshold. Under the fixed threshold classification, if any of the six BISG probabilities of race/ethnicity (Hispanic, Black, API, White, AIAN, and multiracial) is greater than or equal to the threshold, the corresponding race/ethnicity indicator takes the value of 1; otherwise it takes the value of 0. Adjaye-Gbewonyo et al. (2014) find that cutoffs in the range of 0.50–0.57 optimize sensitivity and specificity for White, Black, API, and Hispanic health plan members. Using mortgage and auto loan data, Baines and Courchane (2014) evaluate classification errors of BISG under the 50% and

80% threshold and strongly urge using a threshold no smaller than 50% in evaluating disparities and calculating consumer harm.

Several limitations exist for the fixed threshold classification. First, there are situations in which none of the six BISG probabilities meet the threshold, and therefore the borrower will not be identified as belonging to any of the six race/ethnicity categories, especially when the threshold is set at a high level. We term this situation as "uncovered". On the other side, if the cutoff is less than 50%, a borrower might be "over-covered" with more than one race/ethnicity assigned. Moreover, the distribution of the underlying continuous probability greatly impacts the choice of the threshold. Therefore, one threshold that proves to be optimal on one data sample might no longer be so on another. Generally speaking, the false positives decrease and the false negatives increase as the threshold increases. However, the sensitivity of such changes is determined by the probability distribution. Figure 3 plots the kernel densities of the BISG probabilities, which obviously do not follow a uniform distribution. The observations of Black and API are heavily skewed to the left, where the probability value is low. The density curves for Hispanic and White are bimodal, with high density occurring at the low and high ends of the probability line. For all the race/ethnicity groups, at the probability range 0.2–0.8, the kernel densities are flat and at a very low level, indicating that moving the threshold in this range is unlikely to lead to a significant change in classifying race/ethnicity. This paper presents the BISG fixed with 80% as the threshold (the BISG 80%).[19]

The BISG type of models have been used not only in health studies but also in the machine learning field, under the name of naive Bayes (or simple Bayes, independence Bayes) models because they rely on the conditional independence assumption to simplify the calculation of the posterior probability. A common decision rule used in machine learning research is to pick the most likely classification outcome, which is known as the "maximum a posteriori" or MAP decision rule. Despite its seemingly strong assumption of conditional independence which is often hard to satisfy in reality, naive Bayes classifiers have proved to work quite well. Domingos and Pazzani (1997) provide a list of such studies and suggest that even though naive Bayes might produce an inaccurate probability estimate for each class, it tends to assign the highest probability to the correct class. Zhang (2004) provides further explanation for the optimality of naive Bayes. His rationale is that it is not just the dependencies between attributes that matter, rather, how they are distributed. In this paper, we leverage the relevant findings in machine learning and propose the max classification approach to dichotomize the BISG proxy. Under this approach, a race/ethnicity indicator takes the value of 1 if the corresponding race/ethnicity prediction is the maximum of the six predictions. The max classification has a full coverage in the sense that it

[19] We explore the BISG fixed with 50% cutoff as well. The conclusions remain unchanged.

guarantees an assigned race/ethnicity for each borrower, therefore eliminating the "uncovered" situation under the fixed threshold classification. It is rare to have "over-covered" either under the BISG max since BISG probabilities are likely to be different across the six race/ethnicity groups. In addition, without a fixed threshold, the race/ethnicity assignment is less dependent on the value of probability whose distribution could vary across data samples.

Panel A of Table 3 tabulates the coverage of the BISG max and BISG 80%. Judging by the number of borrowers with an assigned race/ethnicity, the BISG max has a coverage very similar to that of the self-reported race/ethnicity. The BISG 80% only covers 462,736 borrowers, about 74% of the 627,638 total borrowers for Hispanic, Black, API, and White. The under-coverage of the BISG 80% is more severe for Black, API, and White than for Hispanic.

The binary indicators derived using the two classification approaches are then compared with the reported race/ethnicity. Four outcomes can be created by overlaying the reported over the dichotomized race/ethnicity: No/No (true negatives, TN), No/Yes (false positives, FP), Yes/No (false negatives, FN), and Yes/Yes (true positives, TP). As the BISG 80% demands a high value of probability to assign 1 to a race/ethnicity, we expect it to have lower false positives and higher false negatives than the BISG max.[20] Since the six race/ethnicity probabilities add up to 100%, a probability greater than or equal to 80% must be the maximum probability; however the maximum probability does not necessarily meet the 80% threshold criteria. Panel B of Table 3 tabulates the four outcomes for the BISG max and BISG 80% for each race/ethnicity and the corresponding false positive rate (FPR=FP/(FP+TN)) or Type I error rate and the false negative rate (FNR=FN/(TP+FN)) or Type II error rate. The results support our hypothesis. Taking Hispanic for example, the FPR is 8% for the max classification vs. 4% for the BISG 80%; the max classification FNR is 12%, lower than the 19% FNR for the BISG 80%. It is also noted that the absolute level of FNR for the BISG 80% is quite high for Black, API, and White, taking values of 51%, 48% and 36%, respectively.

Panel C of Table 3 lists the average values of mortgage price (note rate) and important variables in mortgage pricing decisions, including income, FICO, CLTV, and LTV, for each classification outcome (TN, FP, FN, and TP). The first thing observed is that note rate, income, FICO, and CLTV are significantly correlated with the classification outcome groups. Comparison of TN with TP shows that true Hispanic and Black on average have higher note rate and CLTV, and lower income and FICO, than the true non-Hispanic and non-Black, respectively; API and White are the opposite, with lower note rate and CLTV, and higher income and FICO. The mean statistics for the two misclassified groups (FP and

[20] This feature of the BISG 80% makes it potentially more attractive for certain tasks where a lower FPR is valued more than a lower FNR, for instance, to identify consumers who are eligible for restitution or other remediation actions. However, in general, fair lending risks analysis treats FPR and FNR equally.

FN) are less extreme than for the TN and TP groups. Taking the BISG max classification for Hispanic for example, while the TP group has the highest note rate of 7.77% and TN has the lowest note rate of 6.95%, the FP and FN have a note rate in between, at 7.33% and 7.39% respectively. When the race/ethnicity classification errors interact with target and control variables in the regression, it is therefore expected that the disparity estimates will be impacted.

VI. Using the BISG Proxy in Assessing Pricing Disparities in Mortgage Lending

In this section, we evaluate the effect of using BISG proxy in fair lending regression analysis of mortgage pricing. Specifically we compare the proposed BISG max with the BISG 80% and BISG continuous. The pricing disparity estimate generated by the reported race/ethnicity is used as the benchmark to assess the estimation bias.

VI.A. Measuring Pricing Disparities in Mortgage Lending

Since Munnell et al. (1996) introduced multivariate regression into fair lending analysis in their famous Boston Fed study, this technique has been widely used in evaluating potential disparities in mortgage lending. Extensive literature[21] exists in testing whether substantial differences exist in mortgage pricing between similarly situated PBGs and CGs. Similarly, we estimate the pricing disparities using ordinary least square (OLS) regression with the specification

$$Note\ Rate = \beta_0 + \beta_r r + \beta_X X + \varepsilon. \quad (7)$$

The rate spread reported by HMDA is less informative than the continuous note rate, as the former is only reportable above a certain threshold.[22] We obtain the continuous note rate from the DataQuick ABS/MBS to allow better capturing of mortgage price differences. The variable *r* identifies the race/ethnicity of the PBG (Hispanic, Black, API, AIAN, or multiracial) from the CG (White). The corresponding coefficient β_r flags potential pricing disparities if it is different from 0 with statistical significance. A significant and positive/negative β_r suggests that the PBG receives a higher/lower price than the CG.

The control vector is denoted as *X*. When *X* is empty, β_r estimates the raw disparity; when *X* contains legitimate factors considered in pricing, β_r estimates the adjusted disparity.[23] We control for common creditworthiness factors and loan characteristics, which include income (in the natural logarithm

[21] For example, Courchane and Nickerson, 1997; Crawford and Rosenblatt, 1999; Black et al., 2003; Boehm et al., 2006; Boehm and Schlottmann, 2007; Courchane, 2007; Bocian et al., 2008; and Zhang, 2013.
[22] Rate spread is the difference between the loan's annual percentage rate (APR) and a survey-based estimate of APRs currently offered on prime mortgage loans of a comparable type. A lender reports rate spread in HMDA if it is equal to or greater than 1.5 percentage points for a first-lien loan or 3.5 percentage points for a subordinate-lien loan.
[23] Because we don't know the full details of the pricing policies and procedures, and the data potentially lack certain factors considered, the disparities estimated in this paper should not be considered as indicative of existence or non-existence of disparities in mortgage pricing decisions.

form), FICO, LTV, and CLTV. Mortgage interest rate tends to differ by collateral type and product type; we therefore control for them in the regression as well. The DataQuick ABS/MBS database identifies collateral types, with the four major ones being jumbo A, Alt A, subprime, and second lien; Eqn. (7) thus includes dummies for the four collateral types with the remaining collateral types grouped as the reference. The DataQuick ABS/MBS product type provides detailed information on product features (such as fixed rate [FRM] vs. adjustable rate [ARM], interest only [IO], or balloon) and loan term for each corresponding product feature. Indicators are created for the top 14 products that comprise 90% of the loans. Certain lenders specialize in specific products or target certain borrowers, which we account for by including a subprime lender indicator[24] in the regression. Furthermore, Eqn. (7) controls for fixed effect of lenders to account for individual lender differences within lender type (subprime vs. non-subprime). Dummies are created for the top 20 lenders that originated about 60% of the loans with the reference group containing the remaining miscellaneous lenders. The specification also includes fixed effects of MSA, loan origination year, and MSA×origination year. Appendix 2 of the e-companion provides more detail of the control variables.

We run regressions of note rate for raw and adjusted disparities using race/ethnicity under the reported, BISG max, BISG 80%, and BISG continuous. With the reported race/ethnicity, the variable *r* is an indicator I_r that takes the value of 1 if a borrower belongs to the PBG and 0 if he/she belongs to the CG. Under the BISG max, $r = 1$ if $p_r = \max_i p_i$ and $r = 0$ if $p_{White} = \max_i p_i$, with *i* being Hispanic, Black, API, AIAN, multiracial, and White. Under the BISG 80%, $r = 1$ if $p_r \geq 80\%$ and $r = 0$ if $p_{White} \geq 80\%$.The regression compares all five PBGs simultaneously against the CG if the BISG continuous is used, so $r = (p_{Hispanic}, p_{Black}, p_{API}, p_{AIAN}, p_{multiracial})$, which is a vector of continuous probabilities for all PBGs. There is an additional regression "Reported II" for adjusted disparity,[25] which is introduced to help decompose the bias of using BISG continuous in estimating pricing disparities. Reported II estimates all reported PBG effects simultaneously with *r* being a vector of the reported race/ethnicity dummies, i.e., $(I_{Hispanic}, I_{Black}, I_{API}, I_{AIAN}, I_{multiracial})$.

Table 4 Panel A tabulates the modeling sample composition for each regression. Separate regression is performed for each PBG under the reported, BISG max, and BISG 80%, so each sample consists of borrowers associated with the particular PBG and the CG. Due to misclassification, the data samples of BISG max and 80% are not exactly the same as those of the reported, but the samples of the

[24] For the period 2004–2007, HUD produced subprime lender lists for 2004 and 2005 (http://www.huduser.gov/portal/datasets/manu.html). The 2004 and 2005 lists are very similar, so we decided to use the 2005 list to create the subprime lender indicator.
[25] Since raw pricing disparity does not consider other control variables except race/ethnicity, the regression of reported II resolves to the reported for raw pricing disparity.

BISG max are much closer. One regression is fitted on the entire sample for all five PBGs against the CG under the reported II and BISG continuous. The count of the PBGs and CG for the BISG continuous is calculated as the weighted average of the corresponding BISG probability. Again it shows that the BISG continuous is a biased estimate of the reported race/ethnicity.

Table 4 then lists the estimated raw and adjusted pricing disparities in Panels B and C, respectively. The coefficient β_r, its standard error, and adjusted R-squared of each regression are provided. We observe that introducing control factors greatly reduces the estimated pricing disparities and improves the goodness of fit, indicating that much of the price differences can be explained by legitimate factors considered in the lender's decision process. For example, based on the reported race/ethnicity, the raw disparity for Hispanic is 1.0143, suggesting that Hispanic paid a note rate 1 percentage point higher than that of White. However, the adjusted price difference is reduced to 11 basis points once considering for income, FICO, LTV, CLTV, collateral, product, lender, MSA, and origination year. Compared with similarly situated White, Black has the largest price difference of about 26 basis points, followed by Hispanic with a price difference of 11 basis points; API has the least price difference, of approximately 3 basis points.

The disparity estimates of the two BISG classifications are much closer to those of the reported than the BISG continuous, presenting a case that dichotomizing a mismeasured covariate reduces bias. The BISG continuous greatly (more than 100 percent) overestimates the pricing disparities for Hispanic, Black, and API, which we attribute to the measurement errors of the continuous probabilities and to the one regression approach that does not allow other covariates to vary by minority group. Comparing estimates under the reported with reported II shows that the impact of performing one regression instead of separate regressions is relatively small: Disparity estimate increases by 1–3 basis points for Hispanic, Black, and API. Therefore, most of the estimation bias displayed by the BISG continuous can be attributed to its measurement errors. For example, if the biased probability instead of reported race/ethnicity is used, ceteris paribus, the disparity estimate for Hispanic increases by 11 basis points from 12 basis points under the reported II to 23 basis points under the BISG continuous.

The BISG max produces disparity estimates closer to those of the reported with smaller standard errors than the BISG 80%, attesting that maximizing posterior probability is the optimal Bayesian decision rule. However, while the BISG max assesses the disparities quite accurately for Hispanic and API, its performance for Black is not as accurate as that for Hispanic and API.

VI.B. Robustness Analyses

We perform a series of analyses to ensure the robustness of the results.

VI.B.1. Bootstrap

We replicate the adjusted pricing disparity analysis using the bootstrap method (Efron, 1979). The HMDA-DQ dataset is 100% sampled with replacement for 1,000 times, then the regressions are run for each replicated sample, resulting in 1,000 regressions and 1,000 β_rs for each scenario. We report the summary statistics (mean and standard error) of the coefficient β_r in Table 5. The mean and standard error of β_r based on the bootstrapped samples are highly consistent with the results contained in Panel C of Table 4.

VI.B.2. Performance by Subsample

We examine the performance of race/ethnicity proxies on subsamples to see if the proxies can capture pricing disparities accurately as they vary across subpopulations. The results are provided in Table 6.

As a prohibited basis, the borrower's gender might be associated with differential pricing practice; therefore we compare the price received by single female vs. by single male applicants. The HMDA-DQ data contain about 172,000 single female applicants and 255,000 single male applicants, with a ratio of 40:60. Based on the reported race/ethnicity, Panel A of Table 6 shows that single female applicants receive a slightly higher, thus unfavorable, price than the single male applicants for API, but not for Hispanic and Black, after accounting for borrower's creditworthiness and loan characteristics. Despite the variations in pricing disparities, the BISG max consistently produces more accurate estimates of pricing disparities than the BISG 80% and BISG continuous.

The same pattern is found when we compare the performance of BISG proxies on other subsamples. Panels B, C, D, and E of Table 6 tabulate the proxy performance by origination year, MSA, collateral type, and lender type, respectively. As the race/ethnicity composition differ by origination year and MSA (as shown in Table 1), or by types of collateral and lender (as shown in Table A2 in the e-companion), the price disparities might change as well. For example, Panel B of Table 6 shows that mortgage price received by Hispanic is 17 basis points higher than similarly situated White in 2004, but the difference decreases to 10, 4, and 11 basis points in 2005, 2006, and 2007, respectively. Among the 10 MSAs, the adjusted price difference between Black and White ranges from 44 basis points in Atlanta to 10 basis points in Boston. The pricing disparities vary by the underlying collateral, with second lien having the highest disparities. Among the first lien loans, Alt A loans show a higher adjusted price difference between a minority race/ethnicity and the White than jumbo A and subprime loans. This is possibly because Alt A loans are usually originated without full documentation and therefore more vulnerable to potential manipulation. Started with a higher raw price, after accounting for the loan and borrower characteristics, loans originated by subprime lenders actually have a smaller price difference for

Hispanic and Black as against White than non-subprime lenders. Despite the variations in price differences displayed by the subsamples, the BISG max consistently generates price disparity estimates with the smallest error among the three proxies, while the BISG continuous significantly over-predicts pricing disparities across subsamples.

VI.B.3. Quantile Regression

The existence of outliers might impact the performance of race/ethnicity proxies differently. To quantify the potential impact of outliers, we replace the OLS regression with quantile regression (Koenker and Bassett, 1978) in estimating the adjusted pricing disparities; specifically, we conduct the median regression (quantile=0.5). As Table 7 reveals, the pricing disparities becomes smaller if estimated using quantile regression, however the BISG max consistently delivers better performance than the other two proxies.

VI.B.4. Price Disparities Using Rate Spread

So far we have used the continuous note rate to measure the price differences. However, note rate is not publicly available as HMDA only reports the rate spread for mortgage pricing, which is a continuous variable truncated above certain thresholds. We compare the proxies using rate spread as another robustness analysis. To identify whether a borrower receives a highly priced mortgage loan or not, we generate a high rate spread indicator that equals 1 if the rate spread is above the threshold and 0 if below. The probability of the high rate spread incidence, $\pi = prob(High\ Rate\ Spread\ Indicator = 1)$, can be modeled using the logistic regression

$$\pi = \frac{1}{1+exp(\beta_0+\beta_r r+\beta_X X)^{-1}}, \quad (8)$$

where the logit, $(\beta_0 + \beta_r r + \beta_X X)$, is essentially the right hand side of Eqn. (7). The variable of interest in the rate spread regression is the adjusted odds ratio for race/ethnicity, $\exp(\beta_r)$. For example, Table 8 reports an adjusted odds ratio of 1.6807 if reported Hispanic is used, meaning that Hispanic is 1.6807 times more likely to have a highly priced mortgage than the CG White. Overall, the BISG max achieves the best performance among the proxies. However, the BISG 80% and BISG continuous predict API slightly better than the BISG max for the discrete outcome.

VII. Proxy Performance for Non-Mortgage Credit Products

Due to the lack of reported race/ethnicity for non-mortgage credits, our analysis so far is limited to mortgage loans. To gain insights on the performance of race/ethnicity on non-mortgage products, we leverage credit bureau data. Credit bureau data is borrower based, thus providing a common basis for mortgages to be linked to non-mortgages belonging to the same borrower. If a mortgage loan in the

HMDA-DQ data can be matched to a mortgage loan in the credit bureau data, a comprehensive view of a borrower's credit profile can be obtained. Since the HMDA-DQ data now have both reported and imputed race/ethnicity for the borrower, by matching it to bureau data, we can then assess the accuracy of race/ethnicity proxies for various non-mortgage products, conditional on the borrower having had a mortgage loan at least once.

We extract borrower's credit bureau information from the OCC's CBD. The CBD is longitudinal, containing a 0.7% random sample of all Equifax credit files at the base year (calendar year of 2005), with new files added each year to rebalance the sample due to attrition. Information archived as of June 30 of each year is provided. The CBD contains five segments, covering consumer, tradeline, collections, inquiries, and public records, and this paper utilizes the first two segments. The consumer segment provides characteristics of a consumer, including a unique consumer identity key (CID), the age and gender, credit score, the archive year, credit record starting date, etc. The tradeline segment lists the detail of a credit account, including consumer and tradeline identity key, account description (account ownership, type of creditor, type of account, loan purpose, lender subscriber code), credit limit or the highest balance, term duration, current balance, payment performance (for the past 48 months and current month), and account dates (for example, open date, report date, close date). The consumer segment and the tradeline segment can be linked by the CID. Types of credit account include mortgage, credit card, auto loan/lease, student loan, consumer loan, small business loan, etc.

The population with mortgages represents a great portion of borrowers with credit products. Figure 4 shows the borrower composition by main credit products (here we consider mortgage, credit card, auto loan, and student loan). Out of the 1.9 million borrowers with 5.3 million transactions reported by CBD during the period 2005–2012, about 38% of them have at least one mortgage. A majority of the borrowers with mortgage have other credit products: 20% CBD borrowers also have credit card and auto loan; 9% have credit card; and 6% have credit card, auto loan, and student loan. The proportion of borrowers with only mortgage is very small (roughly 1%).

We carefully match the HMDA-DQ data with CBD mortgage trades of the period 2005–2012 by origination date, location of the property, loan amount, and loan term. Out of the 630,807 HMDA-DQ loans, 6,648 (1.06%)[26] loans find a match in CBD. Then through the CID, we extract borrower's non-mortgage (including credit card, auto loan, and student loan) tradelines. The tri-merged data is termed HMDA-DQ-CBD.[27] Its distribution of borrowers is very similar to that of the mortgage subset of CBD (as

[26] The matching rate of 1.06% is reasonable, given that CBD is a 0.7% random sample of Equifax data.
[27] Note that the non-mortgage trades can be originated before or after the matched mortgage trade, as long as they have ever been reported to CBD during 2005–2012, which enables us to obtain more non-mortgage tradelines.

shown in Figure 5), suggesting that matching with PLS mortgages does not distort the population, which represents the universe of mortgage borrowers.

Panel A of Table 9[28] tabulates the race/ethnicity composition of the HMDA-DQ-CBD data by product. Compared with the HMDA-DQ data (as shown in Panel A of Table 1), the mortgage trades have very similar race/ethnicity composition, again suggesting the good quality of matching HMDA-DQ with CBD. Conditional on having a mortgage loan, borrowers on average have 3.6 credit cards, 1.4 auto loans, and 0.42 student loans. The race/ethnicity of credit card and auto loan remain similar to that of mortgage, but a significant shift is observed in student loan. The proportion of Hispanic decreases from 30.4% to 25.3% and the proportion of Black increases from 10.6% to 16.7%. Overall the distribution of race/ethnicity is intuitive given the product.

The BISG proxy is compared with the geo, surname, and geo-surname proxies in predicting the reported race/ethnicity for each product. Panel B of Table 9 lists the Pearson correlation coefficient between the proxy and the reported race/ethnicity. The BISG proxy is shown to have the highest correlation coefficient, followed by geo-surname proxy. The geo and surname proxies that use a single data source are less correlated with the reported race/ethnicity. The surname proxy performs better than the geo proxy for Hispanic, API, and White, but not for Black. The proxies demonstrate the same pattern as shown by the AUC statistics reported in Panel C of Table 9. We also evaluate the classification approaches using the HMDA-DQ-CBD data. As Panel D of Table 9 shows, the BISG max replicates the reported race/ethnicity fairly well, while the BISG 80% cannot assign a definitive race/ethnicity to almost 30% of the population for each credit product.

The bureau data do not report either underwriting or pricing outcome, therefore we cannot evaluate proxies in estimating disparities in non-mortgage lending. Despite the limitation of the non-mortgage analysis data, it is reassuring to see that the univariate results on mortgage are replicable on non-mortgage credits.

VIII. Conclusion and Future Developments

Often, as race/ethnicity is not available for non-mortgage products, conducting fair lending analysis involves imputing race/ethnicity first. One common remedy is to proxy for race/ethnicity using publicly available information provided by the Census Bureau on geographic and/or surname composition. Compared with the geo-only or surname-only proxy, hybrid approaches are proposed, mainly in other fields, to combine both surname and geographic information to improve accuracy, including the simple linear rule or the more sophisticated BISG proxy. The CFPB has adapted the use of

[28] Table 9 reports the bureau analysis based on transactions. Results based on borrowers are very similar and therefore are not tabulated for brevity.

the BISG methodology in regulating indirect auto lending, as showcased by the Ally Bank case. However, research on assessing fair lending risks using race/ethnicity proxies is still limited, and this paper aims to fill the void.

In addition to the geo or surname proxy which uses a single data source, we evaluate the hybrid BISG and the geo-surname proxy which is constructed using the linear rule. Based on mortgage data, we gauge the performance of the four proxies in predicting the reported race/ethnicity by prediction bias, correlation coefficient, and discriminatory power, and find that the BISG is the best predictor. Specifically, our analysis shows that considering both surname and geo instead of just one source of information significantly improves accuracy, and the BISG performs better than the geo-surname proxy. Furthermore, we evaluate the impact of using race/ethnicity proxies on fair lending assessment. Besides the BISG continuous and BISG fixed used in existing analyses, we introduce the BISG max, which assigns a borrower to the race/ethnicity with the maximum probability. Pricing disparities estimated using the three BISG proxies are compared with those using the reported race/ethnicity. The BISG max and BISG 80% produce more accurate pricing disparity estimates than the BISG continuous. Between the two BISG classifiers, the BISG max surpasses the BISG 80% as expected. The above conclusions withstand comprehensive robustness tests. We extend the univariate analysis to a non-mortgage dataset and again find the superiority of the BISG proxy in predicting race/ethnicity and the better coverage of the BISG max than BISG fixed.

While our analysis shows that the BISG method is more accurate in approximating race/ethnicity than other methods, it is important to recognize that it is still a proxy method and inevitably it has measurement errors compared with the reported race/ethnicity. Although we have shown that dichotomizing the imprecise race/ethnicity could reduce the impact of measurement errors in assessing disparities, as fair lending risks of non-mortgage credit products gain more and more attention, regulators might want to consider the ultimate solution, which is to require lenders to collect race/ethnicity and gender information for non-HMDA products as well.

In the interim, we ought to continually understand and improve the proxies. Besides refining the proxy algorithm, another direction to improve the proxies is to leverage additional information that is indicative of race/ethnicity. Our study, as well as existing research (Elliott et al., 2008 and 2009), has shown that proxies utilizing both geographic and surname surpass those single-sourced proxies. One promising data is the first name, for which currently there is no Census database, as there is for surname or geography. Nanchahal et al. (2001) develop an algorithm, the South Asian Names and Group Recognition Algorithm (SANGRA), to identify South Asian ethnicity based on surname and first name, surname only, first name only, or middle name only. Coldman et al. (1988) consider first and middle

names in addition to surname in identifying Chinese. A publicly available database on first name can facilitate more comprehensive assessment of using first name in proxying for race/ethnicity.

Acknowledgements

The author is grateful to the department editor (Amit Seru), an associate editor, and two anonymous reviewers for their constructive comments. I thank Marc Elliott for sharing the BISG algorithm; Tom Mayock for providing the HMDA-DataQuick merged data that are used to derive my analysis data; and Mike Carhill, Chau Do, Yingyao Hu, Marilyn Jacob, Phillip Li, Tom Mayock, and Xinlei Zhao for helpful advice and comments. The views expressed in this paper are those of the author and do not necessarily reflect the views of the Office of the Comptroller of the Currency or the U.S. Department of the Treasury.

References

Adjaye-Gbewonyo, D., R. A. Bednarczyk, R. L. Davis, and S. B. Omer. 2014. Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study. *Health services research* 49 (1), 268–283.

Baines, A. P. and M. J. Courchane. 2014. Fair Lending: Implications for the Indirect Auto Finance Market. Study prepared for the American Financial Services Association, available online at http://www.crai.com/sites/default/files/publications/Fair-Lending-Implications-for-the-Indirect-Auto-Finance-Market.pdf

Black, H., T. Boehm, and R. DeGennaro. 2003. Is There Discrimination in Mortgage Pricing? The Case of Overages. *Journal of Banking and Finance* 27: 1139–1165.

Bocian, D., K. Ernst, and W. Li. 2008. Race, Ethnicity and Subprime Home Loan Pricing. *Journal of Economics and Business* 60: 110–124.

Boehm, T. and A. Schlottmann. 2007. Mortgage Pricing Differentials Across Hispanic, African-American, and White Households: Evidence from the American Housing Survey. *Cityscape: A Journal of Policy Development and Research* 9 (2): 93–136.

Boehm, T., P. Thistle, and A. Schlottmann. 2006. Rates and Race: An Analysis of Racial Disparities in Mortgage Rates. *Housing Policy Debate* 17 (1): 109–149.

CFPB White Paper. 2014. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment. Available online at http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

Cohen-Cole, E. 2011. Credit Card Redlining. *Review of Economics and Statistics* 93 (3): 700–713.

Coldman, A. J., T. Braun, and R. P. Gallagher. 1988. The Classification of Ethnic Status Using Name Information. *Journal of Epidemiology and Community Health* 42: 390–395.

Courchane, M. and D. Nickerson. 1997. Discrimination Resulting from Overage Practices. *Journal of Financial Services Research* 11: 133–151.

Courchane, M. J. 2007. The Pricing of Home Mortgage Loans to Minority Borrowers: How Much of the APR Differential Can We Explain? *Journal of Real Estate Research* 29 (4): 399–439.

Crawford, G. and E. Rosenblatt. 1999. Differences in the Cost of Mortgage Credit Implications for Discrimination. *The Journal of Real Estate Finance and Economics* 19: 147–159.

Domingos, P. and M. Pazzani. 1997. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning* 29: 103–130.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7: 1–26.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. 2009. Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Services and Outcomes Research Methodology* 9: 69–83.

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie. 2008. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. *Health Services Research* 43 (5): 1722–1736.

Fiscella, K. and A. M. Fremont. 2006. Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity. *Health Services Research* 41 (4 Pt 1): 1482–1500.

Gustafson, P. and N. D. Le. 2002. Comparing the Effects of Continuous and Discrete Covariate Mismeasurement, with Emphasis on the Dichotomization of Mismeasured Predictors. *Biometrics* 58: 878–887.

Koenker, R. and G. Bassett. 1978. Quantile Regression. *Econometrica* 46 (1): 33–50.

Munnell, A. H., G. M. B. Tootell, L. E. Browne, and J. McEneaney. 1996. Mortgage Lending in Boston: Interpreting HMDA Data. *American Economic Review* 86 (1): 25–53.

Mayock, T. and R. Sprizer. 2015. Socioeconomic and Racial Disparities in the Financial Returns to Homeownership. SSRN working paper: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2595471.

McCaffrey, D. F. and M. N. Elliott. 2008. Power of Tests for a Dichotomous Independent Variable Measured with Error. Health Services Research 43 (3): 1085–1101.

Nanchahal, K., P. Mangtani, M. Alston, and I. dos Santos Silva. 2001. Development and Validation of a Computerized South Asian Names and Group Recognition Algorithm (SANGRA) for Use in British Health-Related Studies. *Journal of Public Health Medicine* 23: 278–85.

Zhang, H. 2004. The Optimality of Naive Bayes. *Proceedings of the 17th International FLAIRS Conference (FLAIRS 2004)*.

Zhang, Y. 2013. Fair Lending Analysis of Mortgage Pricing: Does Underwriting Matter? *The Journal of Real Estate Finance and Economics* 46: 131–151.

Table 1: Reported Race/Ethnicity Composition

Panel A: Overall

Race/Ethnicity	U.S.[a]	10-MSA[a]	10-MSA Adult[a]	HMDA-DQ
Hispanic	17.3%	20.2%	18.6%	33.0%
Black	12.1%	16.3%	15.5%	11.5%
API	4.8%	8.0%	8.2%	9.9%
White	63.0%	53.2%	55.8%	45.1%

Panel B: by MSA

Race/Ethnicity	Atlanta	Boston	Chicago	Detroit	Los Angeles	Miami	New York	Pittsburgh	San Diego	San Francisco
Hispanic	5.0%	11.5%	25.4%	2.7%	41.7%	**52.1%**	23.3%	1.2%	32.3%	25.2%
Black	**45.0%**	8.8%	19.2%	22.0%	5.4%	14.3%	13.1%	9.5%	3.0%	5.7%
API	3.6%	5.3%	5.4%	2.8%	13.6%	1.7%	8.5%	1.3%	10.6%	**24.2%**
White	46.0%	74.1%	49.6%	71.9%	38.7%	31.6%	54.6%	**87.5%**	53.3%	44.2%
% of Obs.	4.1%	4.4%	12.6%	2.0%	27.4%	18.2%	9.8%	0.7%	8.0%	12.8%

Panel C: by Origination Year

Race/Ethnicity	2004	2005	2006	2007
Hispanic	31.2%	32.9%	36.4%	24.0%
Black	9.4%	11.5%	13.5%	9.3%
API	10.1%	10.2%	9.0%	12.3%
White	48.7%	45.0%	40.6%	53.9%
% of Obs.	21.2%	36.8%	33.4%	8.7%

The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge.

[a] Calculated using Census 2010 SF1.

Table 2: Prediction Performance of Race/Ethnicity Proxies

Panel A: Prediction Bias

Race/Ethnicity	BISG	Geocoding	Surname	Geo-Surname
Hispanic	1.6%	-4.2%	-1.8%	-3.0%
Black	-0.0%	2.1%	-2.5%	-0.2%
API	-0.4%	0.2%	-2.4%	-1.1%
White	-2.8%	0.3%	4.8%	2.5%

Panel B: Correlation Coefficient

Race/Ethnicity	BISG	Geocoding	Surname	Geo-Surname
Hispanic	0.83	0.58	0.80	0.81
Black	0.74	0.57	0.54	0.67
API	0.73	0.41	0.68	0.70
White	0.76	0.56	0.66	0.72

This table reports Panel A) the prediction bias of the proxies as the difference between their mean predicted probability and the mean reported race/ethnicity, and Panel B) the Pearson correlation coefficients of the proxies with self-reported race/ethnicity. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. All the correlation coefficients are statistically significant at the 1% significance level.

Table 3: BISG Classification

Panel A: Coverage

Race/Ethnicity	Reported	BISG Max	BISG 80%
Hispanic	207,888	216,939	184,130
Black	72,596	67,387	41,477
API	62,727	51,632	37,324
White	284,427	294,069	199,805
Total	627,638	630,027	462,736

Panel B: Classification Errors

Race/Ethnicity	TN	FP	FN	TP	FPR	FNR
BISG Max						
Hispanic	389,918	33,001	23,950	183,938	8%	12%
Black	541,139	17,072	22,281	50,315	3%	31%
API	557,244	10,836	21,931	40,796	2%	35%
White	297,382	48,998	39,356	245,071	14%	14%
BISG 80%						
Hispanic	406,225	16,694	40,452	167,436	4%	19%
Black	552,053	6,158	37,277	35,319	1%	51%
API	563,443	4,637	30,040	32,687	1%	48%
White	329,069	17,311	101,933	182,494	5%	36%

Panel C: Classification Errors and Borrower/Loan Characteristics

Race/Ethnicity	Variables	BISG Max				BISG 80%			
		TN	FP	FN	TP	TN	FP	FN	TP
Hispanic	Note Rate	6.95	7.33	7.39	7.77	6.96	7.42	7.46	7.79
	Income	163,411	134,987	134,938	112,669	162,627	126,301	134,147	110,665
	FICO	700	690	685	680	700	687	685	680
	CLTV	86	89	89	91	86	90	89	91
	LTV	68	65	65	63	68	65	65	63
Black	Note Rate	7.06	7.84	7.92	8.48	7.07	8.26	8.07	8.57
	Income	153,530	108,549	132,062	84,480	152,763	97,604	117,634	79,506
	FICO	699	677	668	653	698	665	664	651
	CLTV	87	90	91	92	87	91	91	92
	LTV	66	67	65	67	66	68	65	67
API	Note Rate	7.30	6.74	7.03	6.41	7.29	6.63	6.97	6.32
	Income	143,366	171,018	153,319	172,120	143,599	179,665	155,001	175,238
	FICO	691	708	700	719	691	713	702	722
	CLTV	88	87	88	84	88	86	88	84
	LTV	66	68	65	69	66	69	66	69
White	Note Rate	7.69	7.39	7.26	6.62	7.67	7.23	6.91	6.59

	Income	117,106	146,440	138,375	182,319	119,114	161,953	154,154	188,573
	FICO	681	685	692	710	681	689	701	711
	CLTV	90	89	88	85	90	88	87	84
	LTV	64	66	67	69	65	67	68	69

The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. The BISG max classification assigns a borrower to a race/ethnicity if the corresponding probability is the maximum among the six race/ethnicities probabilities; the BISG 80% classification assigns a race/ethnicity if the corresponding probability is greater than or equal to 80%. TN: true negatives, FP: false positives, FN: false negatives, TP: true positives, FPR: false positive rate, FNR: false negative rate.

Table 4: Assessing Pricing Disparities Using BISG Proxies

Panel A: Modeling Samples

PBG	Population	Reported	BISG Max	BISG 80%
Hispanic	PBG	207,888	216,939	184,130
	CG	284,427	294,069	199,805
	Total	**492,315**	**511,008**	**383,935**
Black	PBG	72,596	67,387	41,477
	CG	284,427	294,069	199,805
	Total	**357,023**	**361,456**	**241,282**
API	PBG	62,727	51,632	37,324
	CG	284,427	294,069	199,805
	Total	**347,154**	**345,701**	**237,129**

Population		BISG Continuous	Reported II
PBG	Hispanic	217,964	207,888
	Black	72,883	72,596
	API	60,417	62,727
	AIAN	959	1,596
	Multiracial	12,016	1,573
CG	White	266,568	284,427
Total		630,807	630,807

Panel B: Raw Pricing Disparities

PBG	Statistics	Reported	BISG Max	BISG 80%	BISG continuous
Hispanic	Coefficient	1.0143***	0.9519***	1.1040***	1.2055***
	Std. Error	0.0069	0.0068	0.0076	0.0081
	Adj. R-Squared	0.0420	0.0369	0.0515	0.0704
Black	Coefficient	1.6031***	1.5731***	1.8729***	2.1292***
	Std. Error	0.0096	0.0100	0.0122	0.0125
	Adj. R-Squared	0.0722	0.0647	0.0897	0.0704
API	Coefficient	-0.0810***	-0.2694***	-0.2939***	-0.3117***
	Std. Error	0.0103	0.0112	0.0128	0.0132
	Adj. R-Squared	0.0002	0.0017	0.0022	0.0704

Panel C: Adjusted Pricing Disparities

PBG	Statistics	Reported	BISG Max	BISG 80%	BISG continuous	Reported II
Hispanic	Coefficient	0.1085***	0.1241***	0.1664***	0.2273***	0.1221***
	Std. Error	0.0041	0.0039	0.0048	0.0048	0.0039
	Adj. R-Squared	0.7617	0.7632	0.7634	0.7621	0.7610
Black	Coefficient	0.2584***	0.3320***	0.4513***	0.5250***	0.2840***
	Std. Error	0.0057	0.0059	0.0078	0.0072	0.0055
	Adj. R-Squared	0.7476	0.7493	0.7447	0.7621	0.7610
API	Coefficient	0.0294***	0.0357***	0.0441***	0.1011***	0.0584***
	Std. Error	0.0055	0.0059	0.0069	0.0069	0.0055
	Adj. R-Squared	0.7453	0.7481	0.7401	0.7621	0.7610

This table reports raw and adjusted pricing disparities (in note rate difference) using reported (also reported II for adjusted pricing disparities), BISG max, BISG 80%, and BISG continuous race/ethnicity, estimated by OLS regression. Regression of the adjusted pricing disparities controls for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 5: Adjusted Pricing Disparities: Bootstrap

PBG	Statistics	Reported	BISG Max	BISG 80%	BISG Continuous
Hispanic	Mean	0.1085***	0.1240***	0.1664***	0.2272***
	Std. Error	0.0041	0.0041	0.0051	0.0049
Black	Mean	0.2583***	0.3317***	0.4508***	0.5247***
	Std. Error	0.0063	0.0065	0.0087	0.0079
API	Mean	0.0294***	0.0358***	0.0442***	0.1013***
	Std. Error	0.0053	0.0055	0.0065	0.0063

The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. Then the data is resampled with replacement for 1,000 times. Each sample is estimated for the adjusted pricing disparities (in note rate difference) using reported, BISG max, BISG 80%, and BISG continuous race/ethnicity. The OLS regression is used to estimate the disparities, controlling for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The 1,000 estimated pricing disparity coefficients are then used to calculate the mean and standard error. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 6: Adjusted Pricing Disparities by Subsample

Panel A: by Applicant's Gender

Gender	PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Single Female	Hispanic	0.0862***	0.0996***	0.1374***	0.2012***
	Black	0.2347***	0.2962***	0.3943***	0.4718***
	API	0.0409***	0.0443***	0.0475***	0.1148***
Single Male	Hispanic	0.0883***	0.1092***	0.1584***	0.2124***
	Black	0.2404***	0.3207***	0.4405***	0.5216***
	API	0.0119	0.0218**	0.0318**	0.0838***

Panel B: by Origination Year

Origination Year	PBG	Reported	BISG Max	BISG 80%	BISG Continuous
2004	Hispanic	0.1738***	0.1928***	0.2412***	0.3103***
	Black	0.3244***	0.4097***	0.5614***	0.6477***
	API	0.0206**	0.0189*	0.0168	0.0937***
2005	Hispanic	0.1020***	0.1290***	0.1637***	0.2267***
	Black	0.2412***	0.3150***	0.4244***	0.5099***
	API	0.0438***	0.0511***	0.0646***	0.1096***
2006	Hispanic	0.0360***	0.0461***	0.0812***	0.1347***
	Black	0.1733***	0.2396***	0.3304***	0.3778***
	API	-0.0097	0.0247**	0.0275**	0.0787***
2007	Hispanic	0.1084***	0.1021***	0.1391***	0.1885***
	Black	0.2950***	0.3696***	0.5198***	0.5606***
	API	0.0499***	0.0229	0.0314*	0.0722***

Panel C: by MSA

MSA	PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Atlanta	Hispanic	0.1415***	0.1904***	0.2316***	0.3778***
	Black	0.4407***	0.5060***	0.6648***	0.7282***
	API	0.0811*	0.0861*	0.1165**	0.1031*
Boston	Hispanic	0.0922***	0.1138***	0.1399***	0.1945***
	Black	0.1044***	0.1241***	0.1918***	0.2251***
	API	-0.0020	-0.0239	0.0184	-0.0110
Chicago	Hispanic	0.0603***	0.0985***	0.1340***	0.1843***
	Black	0.2011***	0.3062***	0.3618***	0.4202***
	API	-0.0414**	-0.0106	-0.0213	-0.0423*
Detroit	Hispanic	0.1848**	0.0276	0.1992	0.1450
	Black	0.3173***	0.4998***	0.6245***	0.7027***
	API	0.0078	0.1186	-0.0263	-0.0891
Los Angeles	Hispanic	0.1620***	0.1753***	0.2440***	0.2704***
	Black	0.1797***	0.2380***	0.2987***	0.4108***
	API	0.0733***	0.0711***	0.1009***	0.1389***
Miami	Hispanic	0.0931***	0.0918***	0.1320***	0.1699***
	Black	0.1867***	0.2068***	0.3172***	0.3934***
	API	0.0351	0.0252	-0.0155	0.0271
New York	Hispanic	0.1825***	0.2418***	0.2904***	0.3446***
	Black	0.2181***	0.3282***	0.4043***	0.4605***

	API	-0.0160	0.0114	0.0029	-0.0014
Pittsburgh	Hispanic	-0.1811	-0.0694	NA[a]	-0.4720
	Black	0.1961***	0.1366	0.2354*	0.2544**
	API	-0.0450	-0.1011	0.0027	-0.4822*
San Diego	Hispanic	0.1616***	0.1637***	0.2000***	0.2358***
	Black	0.2248***	0.1809***	0.1790	0.4032***
	API	0.0650***	0.0438**	0.0174	0.0522**
San Francisco	Hispanic	0.1205***	0.1166***	0.1684***	0.1930***
	Black	0.2003***	0.2189***	0.4523***	0.4128***
	API	0.0439***	0.0353***	0.0613***	0.0913***

Panel D: by Collateral Type

Collateral Type	PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Jumbo A	Hispanic	0.0864***	0.0941***	0.1243***	0.1340***
	Black	0.0852***	0.1504***	0.2781***	0.2340***
	API	0.0253***	0.0220***	0.0253***	0.0350***
Alt A	Hispanic	0.0909***	0.1025***	0.1299***	0.1764***
	Black	0.2008***	0.2816***	0.4077***	0.4343***
	API	0.0138**	0.0327***	0.0388***	0.0703***
Subprime	Hispanic	0.0355***	0.0628***	0.0954***	0.1363***
	Black	0.1080***	0.1828***	0.2450***	0.3233***
	API	0.0314***	0.0614***	0.0969***	0.0888***
Second Lien	Hispanic	0.2494***	0.2399***	0.3234***	0.4100***
	Black	0.3348***	0.4111***	0.5526***	0.6565***
	API	0.0745***	0.0760***	0.1030***	0.1304***

Panel E: by Lender Type

Lender Type	PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Subprime	Hispanic	0.0413***	0.0756***	0.1098***	0.1475***
	Black	0.1188***	0.1974***	0.2714***	0.3453***
	API	0.0440***	0.0596***	0.0985***	0.0776***
Non-Subprime	Hispanic	0.1199***	0.1280***	0.1638***	0.2204***
	Black	0.3066***	0.3809***	0.5330***	0.5635***
	API	0.0233***	0.0277***	0.0324***	0.0858***

This table reports adjusted pricing disparities (in note rate difference) using reported, BISG max, BISG 80%, and BISG continuous race/ethnicity by subsample. The OLS regression is used to estimate the disparities, controlling for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

[a] Pittsburgh reports "NA" for Hispanic under BISG 80% classification as the algorithm categorizes no one as Hispanic.

Table 7: Adjusted Pricing Disparities by Quantile Regression

PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Hispanic	0.0778***	0.0890***	0.1217***	0.1683***
Black	0.2308***	0.3039***	0.4374***	0.4861***
API	0.0254***	0.0304***	0.0397***	0.0885***

This table reports adjusted pricing disparities (in note rate difference) using reported, BISG max, BISG 80%, and BISG continuous race/ethnicity. The quantile regression (at median) is used to estimate the disparities, controlling for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 8: Adjusted Pricing Disparities using Rate Spread

PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Hispanic	1.6807***	1.6057***	1.8144***	2.0326***
Black	2.1423***	2.0265***	2.4962***	2.9752***
API	1.2015***	1.1043***	1.1331***	1.1191***

This table reports adjusted pricing disparities (in odds ratio of high rate spread incidence) using reported, BISG max, BISG 80%, and BISG continuous race/ethnicity. The logistics regression is used to estimate the disparities, controlling for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 9: Proxy Performance for Non-Mortgage Products

Panel A: Race/Ethnicity Composition

Race/Ethnicity	Mortgage	Credit Card	Auto Loan	Student Loan
Hispanic	30.4%	28.9%	30.6%	25.3%
Black	10.6%	9.2%	10.9%	16.7%
API	10.1%	10.7%	9.1%	9.6%
White	48.4%	50.6%	49.0%	47.7%
# Obs.	6,648	23,699	9,084	2,806
# Obs./ # Obs. of Mortgage		3.6	1.4	0.42

Panel B: Correlation Coefficient

Product	Race/Ethnicity	BISG	Geocoding	Surname	Geo-Surname
Mortgage	Hispanic	0.83	0.58	0.80	0.81
	Black	0.73	0.57	0.51	0.67
	API	0.74	0.42	0.69	0.70
	White	0.76	0.56	0.66	0.73
Credit Card	Hispanic	0.83	0.57	0.80	0.80
	Black	0.71	0.55	0.49	0.65
	API	0.75	0.42	0.69	0.71
	White	0.76	0.56	0.67	0.73
Auto Loan	Hispanic	0.83	0.56	0.79	0.80
	Black	0.74	0.58	0.51	0.68
	API	0.72	0.40	0.66	0.68
	White	0.76	0.56	0.65	0.73
Student Loan	Hispanic	0.82	0.54	0.78	0.79
	Black	0.74	0.61	0.53	0.70
	API	0.75	0.41	0.68	0.71
	White	0.73	0.53	0.59	0.69

Panel C: AUC

Product	Race/Ethnicity	BISG	Geocoding	Surname	Geo-Surname
Mortgage	Hispanic	0.954	0.846	0.932	0.953
	Black	0.955	0.881	0.885	0.936
	API	0.940	0.826	0.902	0.930
	White	0.924	0.824	0.874	0.919
Credit Card	Hispanic	0.951	0.840	0.928	0.950
	Black	0.953	0.874	0.884	0.932
	API	0.942	0.828	0.904	0.930
	White	0.925	0.821	0.876	0.920
Auto Loan	Hispanic	0.952	0.839	0.929	0.951
	Black	0.954	0.880	0.883	0.937

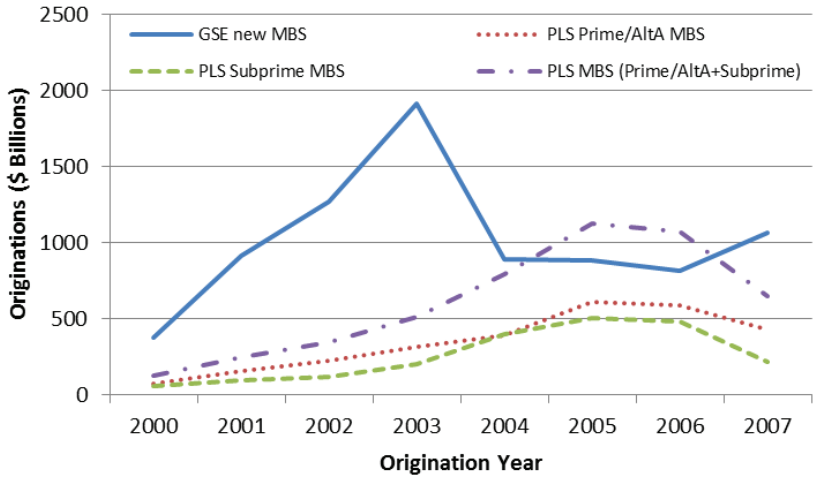
	API	0.935	0.827	0.896	0.926
	White	0.922	0.823	0.869	0.918
Student Loan	Hispanic	0.948	0.836	0.915	0.948
	Black	0.946	0.879	0.870	0.927
	API	0.947	0.844	0.906	0.934
	White	0.907	0.804	0.841	0.901

Panel D: Coverage

Product	Race/Ethnicity	Reported	BISG Max	BISG 80%
Mortgage	Hispanic	2,021	2,114	1,788
	Black	704	663	368
	API	670	584	416
	White	3,215	3,279	2,251
	Total	6,610	6,640	4,823
Credit Card	Hispanic	6,844	7,142	5,972
	Black	2,189	2,026	1,058
	API	2,543	2,252	1,604
	White	11,981	12,251	8,486
	Total	23,557	23,671	17,120
Auto Loan	Hispanic	2,778	2,924	2,462
	Black	988	942	523
	API	831	710	478
	White	4,447	4,497	3,103
	Total	9,044	9,073	6,566
Student Loan	Hispanic	710	729	614
	Black	470	465	252
	API	270	229	173
	White	1,338	1,383	945
	Total	2,788	2,806	1,984

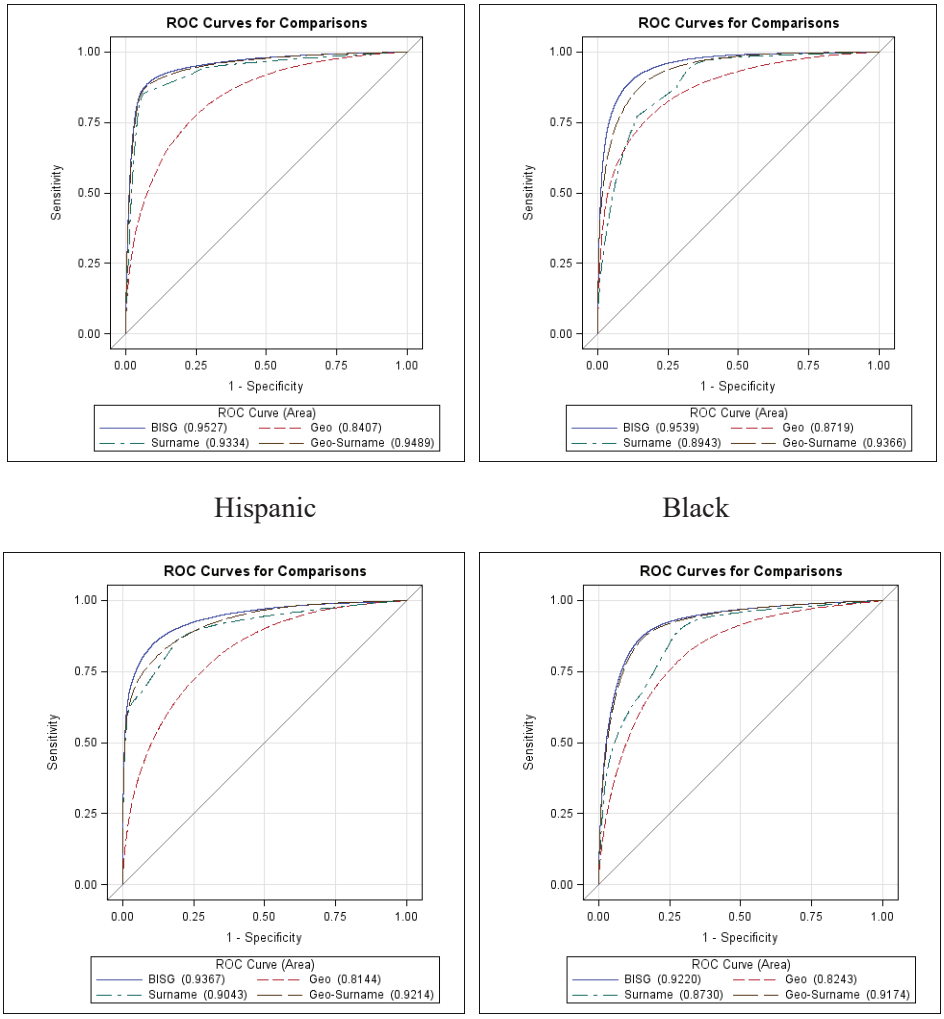
Data source: HMDA-DQ-CBD tri-merged data. All the correlation coefficients are statistically significant at the 1% significance level.

Figure 1: Growth of Private Label Securitized Mortgage Loans



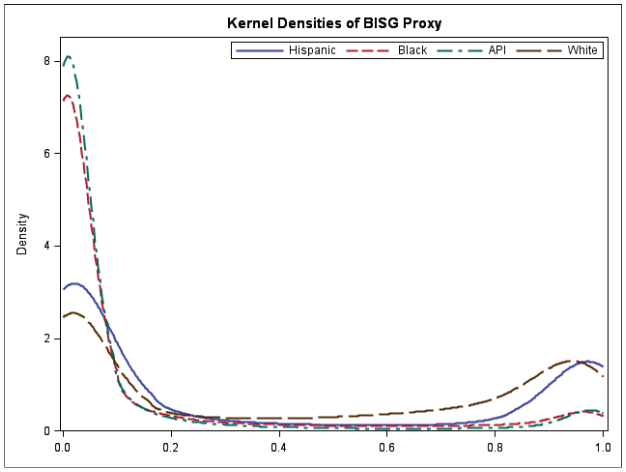
Source: Inside Mortgage Finance (October 31, 2008), 2008 (42): Page 4.

Figure 2: ROC and AUC of Race/Ethnicity Proxies



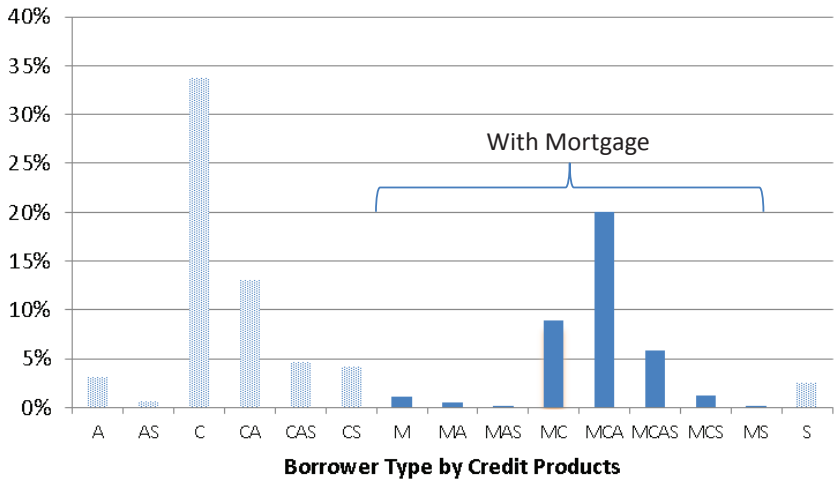
Data source: 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge.

Figure 3: Kernel Densities of BISG Proxy



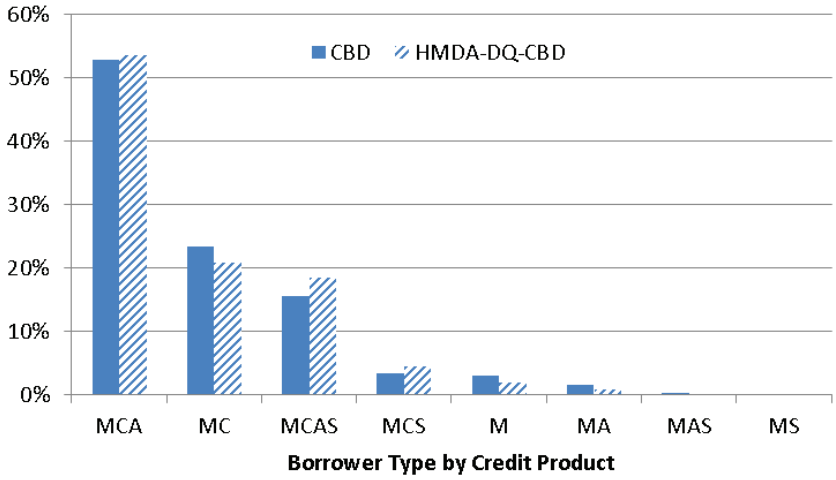
Data source: 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge.

Figure 4: Borrower Composition by Credit Product



Source: OCC CBD database, 2005–2012, covering 1.9 million customers with 5.3 million transactions of mortgage, credit card, auto loan, and student loan. The credit product follows the naming convention of mortgage (M), credit card (C), auto loan (A), and student loan (S).

Figure 5: Borrower Composition Comparison: CBD and HMDA-DQ-CBD

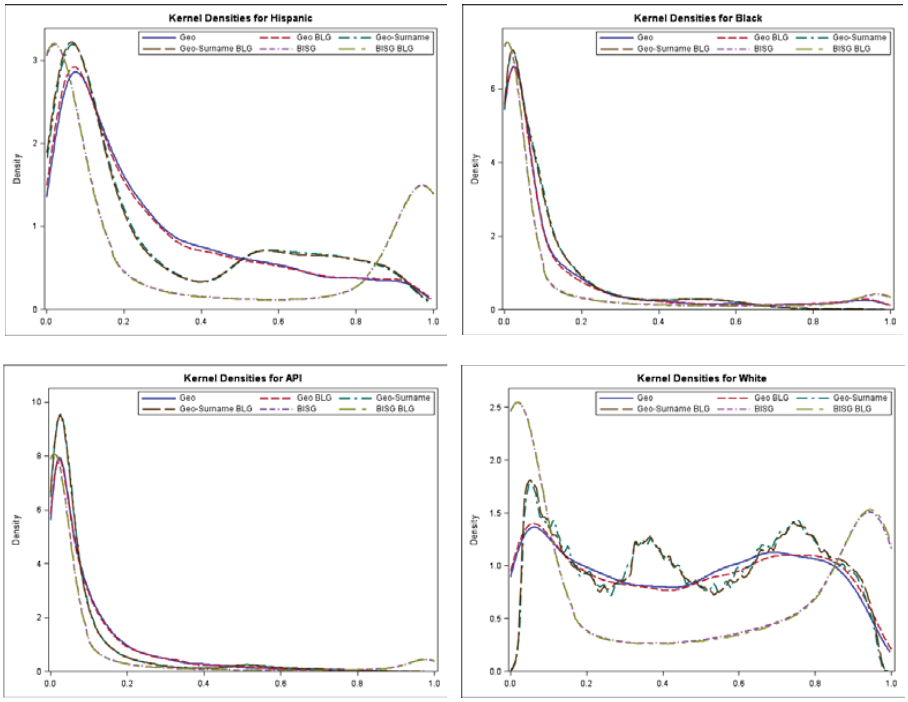


Data source: 1.9 million borrowers from OCC CBD database with mortgage, credit card, auto loan, and student loan tradelines during 2005–2012, and 6,648 borrowers contained in the HMDA-DQ-CBD tri-merged data. The credit product follows the naming convention of mortgage (M), credit card (C), auto loan (A), and student loan (S).

Appendix 1: Compare Race/Ethnicity Proxies Using Census Block Group vs. Tract Level Data

If we use the race/ethnicity composition at the block group instead of tract level, the values of geo, geo-surname, and BISG proxies are subject to change. Out of the total 630,807 loans, 99.35% can be matched to the Census block group level SF1 file; 76 loans have a valid block group but it does not have SF1 population; the remaining 4,037 loans cannot be matched to a valid block group. For the latter two groups with no valid block group SF1 population, the BISG algorithm uses the tract level SF1 file instead. We compare kernel density curves (Figure A1) of the proxies using block group vs. tract level geographic data, and find only minor differences for the geo proxy and no visible differences for the geo-surname and BISG proxies. We further quantify the potential improvement of using the more granular block group level information in assessing price differences. The results are listed in Table A1. Comparing the results using block group level vs. tract level Census data (as shown in Table 4 Panel C), we conclude that there are no significant differences in proxying for race/ethnicity as to which geographic level to use.

Figure A1: Kernel Densities of Proxies using Census Block Group vs. Tract Level Data



Data source: 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. Suffix "BLG" is for proxies calculated using block group Census SF1.

Table A1: Adjusted Pricing Disparities using Block Group Level Census Data

PBG	Reported	BISG Max	BISG 80%	BISG Continuous
Hispanic	0.1085***	0.1278***	0.1695***	0.2300***
Black	0.2584***	0.3373***	0.4535***	0.5260***
API	0.0294***	0.0351***	0.0442***	0.0999***

This table reports adjusted pricing disparities (in note rate difference) using reported, BISG max, BISG 80%, and BISG continuous race/ethnicity using Census block group level data to proxy for race/ethnicity. The OLS regression is used to estimate the disparities, controlling for income (in the form of logarithm), FICO, LTV, CLTV, collateral type, product type, lender type, and fixed effects of lender, MSA, origination year, and MSA×origination year. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Appendix 2: Control Variables Used in Regression Analysis of Mortgage Pricing Disparities

Table A2 provides summary statistics of note rate, income, FICO, LTV, CLTV, and reported race/ethnicities of the entire data sample, as well as subsamples, by types of collateral, product, and lender. It shows that loan pricing, borrower, and loan characteristics do vary significantly by collateral, product, and lender types. For example, compared with the overall population, loans originated with subprime collateral, on average, have a higher note rate, lower income and FICO, higher LTV and CLTV, more Hispanic and Black, and fewer API and White.

Table A2: Summary Statistics

Sample	%	Note Rate	Income	FICO	LTV	CLTV	Hispanic	Black	API
Overall	100%	7.22	146,047	693	66	88	33%	12%	10%
Collateral Type									
Subprime	38.8%	8.22	109,300	654	69	90	44%	18%	7%
Alt A	22.7%	6.41	157,108	717	78	88	28%	8%	10%
Jumbo A	17.1%	5.76	222,760	746	75	79	10%	2%	16%
Second Lien	11.7%	10.35	115,735	684	19	94	43%	13%	8%
Other	9.7%	3.88	169,218	714	73	84	30%	7%	12%
Product Type									
FIXED; 30 Year	18.4%	8.34	151,097	700	53	86	30%	11%	11%
ARM; 2/28	12.1%	7.60	98,751	642	84	90	37%	23%	5%
ARM; 5/25 IO	11.6%	6.07	172,593	725	78	87	24%	6%	13%
BALLOON; 15/30	8.9%	10.44	116,396	680	19	92	31%	12%	10%
ARM; 2/28 IO	8.5%	6.89	117,035	670	82	91	46%	13%	8%
ARM; 30 Year	5.2%	2.58	180,866	712	78	83	51%	11%	10%
FIXED; 30 Year IO	4.9%	6.87	178,943	729	75	87	32%	6%	13%
ARM; 3/27 IO	4.0%	6.10	136,610	706	80	89	22%	7%	9%
ARM; 10/20 IO	3.2%	6.15	232,850	741	75	81	31%	8%	12%
ARM; 3/27	2.9%	7.38	103,578	656	84	89	14%	3%	11%
ARM; 2/28 BALLOON 30/40	2.9%	7.92	121,094	648	82	91	32%	20%	7%
ARM; 7/23 IO	2.6%	6.17	209,969	735	77	84	49%	19%	8%

ARM; 5/25	2.2%	6.19	170,448	717	78	84	17%	4%	12%
ARM; 40 Year	2.0%	4.59	164,726	694	80	89	20%	8%	16%
Other	10.5%	7.58	149,600	697	50	87	40%	11%	12%
Lender Type									
Subprime	30.7%	8.47	108,959	659	61	91	48%	18%	7%
Non-Subprime	69.3%	6.67	162,504	708	69	86	26%	9%	11%

This table reports average values of observables for the overall as well as subsamples by types of collateral, product, and lender. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge.

We also conduct an analysis of variance (ANOVA) to assess the contributions of the controls in Eqn. (7). Table A3 lists the ANOVA Type III F statistics. All the factors are statistically significant with LTV, FICO, CLTV, and collateral type being the top four with the largest F statistics, which is intuitive and supports the setup of the pricing regression analysis.

Table A3: ANOVA Analysis

Source	DF	PBG		
		Hispanic	Black	API
Income (in Log)	1	21***	15***	58***
FICO	1	32,717***	22,845***	19,561***
LTV	1	66,454***	39,769***	37,096***
CLTV	1	16,382***	13,008***	14,737***
Race/Ethnicity	1	715***	2,046***	29***
Collateral Type	4	16,341***	11,252***	11,138***
Product Type	14	7,562***	5,591***	6,117***
Lender Type	1	954***	811***	766***
Lender FE	20	294***	214***	214***
MSA FE	9	844***	623***	503***
Origination Year FE	3	6,036***	6,202***	5,754***
MSA*Origination Year FE	27	78***	59***	57***

This table reports ANOVA Type III F statistics using reported race/ethnicity. The ANOVA is used to evaluate the explanation power of various factors used in Eqn. (7) for note rate. The data contain 630,807 mortgage purchases originated during 2004–2007 in 10 MSAs based on a HMDA and DataQuick merge. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination

Miranda Bogen*
Upturn
mirandabogen@gmail.com

Aaron Rieke
Upturn
aaron@upturn.org

Shazeda Ahmed†
University of California, Berkeley
shazeda@ischool.berkeley.edu

ABSTRACT

Organizations cannot address demographic disparities that they cannot see. Recent research on machine learning and fairness has emphasized that awareness of sensitive attributes, such as race and sex, is critical to the development of interventions. However, on the ground, the existence of these data cannot be taken for granted.

This paper uses the domains of employment, credit, and healthcare in the United States to surface conditions that have shaped the availability of sensitive attribute data. For each domain, we describe how and when private companies collect or infer sensitive attribute data for antidiscrimination purposes. An inconsistent story emerges: Some companies are required by law to collect sensitive attribute data, while others are prohibited from doing so. Still others, in the absence of legal mandates, have determined that collection and imputation of these data are appropriate to address disparities.

This story has important implications for fairness research and its future applications. If companies that mediate access to life opportunities are unable or hesitant to collect or infer sensitive attribute data, then proposed techniques to detect and mitigate bias in machine learning models might never be implemented outside the lab. We conclude that today's legal requirements and corporate practices, while highly inconsistent across domains, offer lessons for how to approach the collection and inference of sensitive data in appropriate circumstances. We urge stakeholders, including machine learning practitioners, to actively help chart a path forward that takes both policy goals and technical needs into account.

ACM Reference Format:
Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3351095.3372877

*Author was affiliated with Upturn at time of writing.
†Author was a Fellow at Upturn at time of writing.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372877

1 INTRODUCTION

Statistical models, including those created with machine learning, can reproduce biases in the historical data used to train them. As powerful institutions increase their reliance upon these models to automate decisions that affect people's rights and life opportunities, researchers have begun developing new techniques to help detect and address these biases. The real-world implementation of these techniques could be an essential part of ensuring the continued viability of civil and human rights protections.

Many machine learning fairness practitioners rely on awareness of sensitive attributes—that is, access to labeled data about people's race, ethnicity, sex, or similar demographic characteristics—to test the efficacy of debiasing techniques or directly implement fairness interventions. A significant body of research presumes the modeler has ready access to data on these characteristics as they build and test their models [13, 19, 21]. The need for this data is plain to see. As a 2003 analysis of racial disparities in healthcare powerfully concluded: "The presence of data on race and ethnicity does not, in and of itself, guarantee any subsequent actions ... to identify disparities or any actions to reduce or eliminate disparities that are found. The absence of data, however, essentially guarantees that none of those actions will occur." [23]

Increasingly, companies that utilize machine learning are being asked to detect and address bias in their products. But they are not the first to grapple with these issues. This paper explores the legal and institutional norms surrounding the collection, inference, and use of sensitive attribute data in three key corporate domains. This analysis has significant implications for machine learning fairness research: If private institutions that mediate access to life opportunities are unable or hesitant to collect or infer sensitive attribute data, then emerging awareness-based techniques to detect and mitigate bias in machine learning models might never be implementable in real-world settings.

Notably, this paper does not discuss complex and important questions about *how* "fairness" should be measured or addressed, recognizing that definitions are manifold [53]. Rather, we make a simpler point: If sensitive attribute data are not available, interventions that rely on them will be severely impaired.

We conduct this exploration through the lens of U.S. civil rights law in the domains of credit, employment, and healthcare. For each domain, we describe when and how private companies collect or infer sensitive attribute data to pursue antidiscrimination goals. These are not the only contexts where collection of sensitive attributes is likely to be justified or important, but they are quintessential areas where such data are already being used to measure and mitigate discrimination. They also highlight major divergences in policy, motivation, and practice.

Comparing these sectors, a complex and inconsistent story emerges. In credit, the law requires some lenders to collect sensitive attribute data, while largely prohibiting others from doing so. In employment, the collection of sensitive attribute data is a familiar part of large employers' day-to-day practice. And in health care, companies' motivation for collecting sensitive attribute data is not just basic antidiscrimination compliance, but rather a moral imperative to address staggering disparities in health outcomes.

We observe that these norms and practices, divergent as they are, typically extend only to traditionally regulated actors. Technology companies that mediate access to opportunities as platforms (e.g., social networks, job boards, and rental sites) or act as vendors to other companies rarely receive clear guidance about when to collect or infer sensitive attribute data. As a result, today, many major technology companies do not collect or infer certain kinds of sensitive attribute data and may therefore struggle to define, detect, and address harms to those protected groups.

We conclude that there are few clear, generally accepted principles about when and why companies should collect sensitive attribute data for antidiscrimination purposes. We emphasize the importance of the machine learning research community engaging on the future development of policy in this area, and urge conversations among stakeholders about whether and how to adapt existing practices or establish new ones.

1.1 Defining "sensitive attribute data"

Throughout this paper, we use the term "sensitive attribute data" to refer to details about people's membership in "protected classes" as defined throughout U.S. civil rights laws. This approach to classification is not without its problems: Rigid categories such as these do not currently accommodate nonbinary identities or membership across multiple groups [3, 34, 37, 51]. We acknowledge the reductive and potentially harmful nature of these classification regimes, while simultaneously emphasizing the importance of understanding how they have motivated data collection practices for bias mitigation, and how the history of these practices can inform contemporary contexts.

1.2 Related work

The fair ML research community has long reflected on the social and policy contexts of its work, recognizing legal tensions [8], historical parallels in prior debates over definitions of fairness [37], and the limitations of data-dependent problem formulation [64, 70]. However, when emphasizing the importance of awareness of sensitive attributes in developing and implementing fairness-enhancing interventions, fair ML research and toolkits [9, 38, 67] often take for granted when framing problems and their solutions that sensitive attribute data are available as inputs [17, 32, 39, 40, 48, 50, 65].

When labeled data are not available, researchers have made powerful discoveries by augmenting existing data through the inference or construction of labeled data de novo [5, 10, 27, 46]. At times, this work has insufficiently acknowledged the full range of challenges to generating or obtaining those data in applied contexts [35, 76, 77].

Veale and Binns [76] and Kilbertus et al [41] propose approaches to dealing with such information deficits without collecting or revealing sensitive data. Chen et al propose a method to impute unavailable protected class data [13]. Veale et al [77] and Holstein et al [35] outline the contextual needs for implementing fairness. Zliobaite and Custers use theoretical and linear regression-based examples to argue that sensitive data must be included in the modeling process in order to avoid discrimination [79]. Focusing on the European regulatory environment, their study distinguishes between direct and indirect discrimination, but does not address how different sectoral laws enable or prevent detection of either type.

This paper aims to bridge gaps between theoretical approaches and practical constraints, extracting lessons for fair ML practitioners from three real-world case studies.

2 CASE STUDIES

2.1 Credit

United States federal law prohibits creditors from discriminating on the basis of certain protected characteristics. However, across the credit sector, there are sharply divergent approaches to collecting sensitive attribute data. On one hand, mortgage lenders are required to collect such data from their borrowers. On the other, consumer lenders are largely prohibited from doing so. The reason for this difference is not immediately apparent, and likely turns on historical details underlying the development of overlapping legal doctrines.

2.1.1 Background. In the mid-1970s, policymakers acknowledged that discriminatory practices in the consumer credit and home mortgage industries shut out women, people of color, low-income groups, and others from accessing these vital economic resources. The Equal Credit Opportunity Act (ECOA), passed in 1974, initially made discrimination on the bases of marital status and sex illegal, and was later expanded to include other protected groups. The following year, the Home Mortgage Disclosure Act (HMDA) similarly made discriminating against low-income home mortgage borrowers illegal. Like ECOA, HMDA grew to encompass categories including race, gender, and national origin through subsequent amendments.

The origins of ECOA trace back to an era when lenders required unmarried women to have male cosigners on their loans. From the outset, regulators feared that mandatory collection of protected class data beyond gender for the purpose of detecting discriminatory lending might itself facilitate such practices. Thus, under ECOA, the collection of these data is banned. Regulation B, which implements ECOA, has made exceptions for some voluntary collection of data on applicants' color, national origin, religion, race, and sex as "monitoring information" in instances where lenders conduct self-testing to determine whether loans are not being granted to individuals on discriminatory grounds.

The Federal Reserve Board (FRB) twice considered amendments to ECOA that would allow voluntary collection of protected class data for non-mortgage loan applicants in order to surface discriminatory lending decisions. In 1995, the first proposal to lift the ban on collecting sensitive attribute data garnered a mix of support and opposition. Noting that discrimination on protected class bases only covered a limited set of criteria for potential disparate treatment during in-person lending scenarios, supporters of the change pointed to the successful identification and reduction of biased

mortgage lending decisions that resulted from HMDA's strict data collection practices [74]. These advocates disagreed with the FRB's long-held claim that recording these data would lead to discrimination in consumer lending, noting that this predicted harm did not unfold in the home mortgage industry.

The argument that voluntary collection of protected characteristic information would lead to discriminatory lending persisted, however, in large part due to credit industry representatives' complaint letters opposing the amendments. Banks and other lending institutions were not inclined to support a measure that would incur higher costs and stricter reporting standards and presumably may have revealed discriminatory practices. They additionally warned that being asked about sensitive attributes could deter some minority applicants. In response to these public comments following the proposed amendment in 1995, the FRB decided to leave the decision about collecting protected class data up to Congress.

After introducing a second proposal to remove the ban on collecting these data in 1998, the FRB once again determined in 2003 that consumer lending institutions should not gather this information. Standing by their original conviction that sensitive attribute information collection would lead to outright discrimination, the FRB also reasoned that making this a voluntary action could result in incomplete data collection and inconsistent data formatting that would hinder cross-market comparison between creditors [74].

The ECOA's evolution was in many ways the opposite of HMDA's expansive push to seek evidence of unfair practices in the mortgage lending industry. HMDA grew out of home mortgage depository institutions' disproportionate withdrawal of investments in largely urban areas from which they drew their deposits: a form of redlining that devitalized older neighborhoods, since residents could not access the credit required to sell and refurbish their homes [42].

HMDA's initial reporting requirements involved publicizing geographic data about lending patterns. As the contexts and causes of home mortgage lending discrimination changed, HMDA was amended between 1980 and the early '00s to expand the scope of institutions covered and to call for reporting of sensitive attribute data on borrowers' gender, race, income, and other categories. When regulators determined these data were insufficient to demonstrate discrimination, they called for further data collection including data about rejected applications and loan pricing.

2.1.2 Data practices. While ECOA prohibits collection of sensitive attribute data for most purposes, its implementing regulations allow banks and "anyone who, in the ordinary course of business, regularly participates in decisions about whether or not to extend credit or how much credit to extend" to collect, in a narrow set of circumstances, sensitive attribute data on individuals applying for non-mortgage loans [71]. If lenders opt to collect this data, they must indicate that the information is being recorded for self-testing and monitoring purposes. If an applicant prefers not to provide their race and sex information, the lender is allowed to make their own determinations of these characteristics from visual observation and surname analysis. If the self-test demonstrates that the institution may have violated ECOA, the lender must attempt to identify the cause and extent of the violation. Save for in some instances, the results of the self-test are considered privileged information that government agencies cannot access in investigations related to ECOA transgressions.

HMDA, by contrast, requires expansive collection of both sensitive attribute data and related mortgage loan application data that can be used to build arguments that discrimination has occurred. Under HMDA, protected data that must be collected as part of a Loan/Application Register (LAR) include sex, race, and ethnicity, with additional requirements that data on income, loan amount and type, property location, and reasons for loan denial (among others) must be reported [42]. Lenders are allowed to use visual observation and surname analysis to guess the sex, race, and ethnicity of applicants who choose not to self-identify these traits. The data are published in different formats depending on the intended recipient. Lenders submit these data to the FRB annually, whereas if a member of the public requested access they would be presented with a modified LAR scrubbed of any identifying information. Finally, the Federal Financial Institutions Examination Council (FFIEC) creates disclosure statements for each lender based on their LAR data, and publishes openly available aggregate reports of HMDA data at city, national, and census-tract levels.

2.1.3 Results and reactions. The question of whether sensitive attribute data should be collected to detect discrimination in consumer lending remains controversial. As one scholar put it, "Even if computerized credit scoring arguably has the potential to eliminate disparate treatment results, disparate impact discrimination may still occur" [74]. Another scholar has suggested creditors should be *required* to conduct self-testing using sensitive attribute data.[4] Lenders and other proponents of credit scoring systems may argue that expanded collection of data on race and other protected class characteristics would be insufficient to prove discrimination given the increasing complexity of how credit scores are calculated.

Today, as was the case when ECOA was passed, the absence of sensitive attribute data makes it difficult to document and mitigate inequitable consumer lending practices. For example, one of the few, robust public studies on credit scores and discrimination in the United States was performed by the FRB in 2007, at the direction of Congress [56]. To conduct its analysis, the FRB created a database that, for the first time, combined sensitive attribute data collected by the Social Security Administration (SSA) with a large, nationally representative sample of individuals' credit records. The FRB noted its study was unique in part because of the lack of sensitive attribute data in this domain, and this unusual undertaking would not have been possible without significant governmental time and resources.

The shortage of sensitive attribute data in the consumer lending space also complicates regulatory enforcement. For example, in 2013, the Consumer Financial Protection Bureau (CFPB) and the Department of Justice found that Ally Financial, an auto lending firm, overcharged over 230,000 minority borrowers on their car loans. Two years later, the CFPB required Ally Financial to send checks from its $80 million settlement to customers believed to have unfairly paid higher prices for their loans [66]. Lacking access to data on which exact individuals had overpaid, however, the CFPB instead used a Bayesian Improved Surname Geocoding (BISG) method to predict which customers were likely to be racial minorities, and were therefore more likely to be victims of Ally Financial's allegedly

discriminatory pricing. Although BISG's probabilistic means of using publicly available surnames and geographical information as proxies for race and ethnicity is regarded as among the most advanced technique of its kind [11], it is not without flaws. In the use of BISG during the Ally Financial payout, some white Americans were misidentified as having been overcharged for car loans on a discriminatory basis and received compensatory checks [7]. Had data collection practices in non-mortgage lending included sensitive attributes, such mistakes could have been averted. Moreover, predictive power of these techniques might diminish over time if housing and marital segregation patterns change.

By contrast, the amendments to HMDA that spurred collection of protected class data came into effect in 1990, and data from 1992 reflected a significant rise in mortgage lending to low- and moderate-income and minority communities [49]. Moreover, in the longer term, the publication of the 1991 data fueled community activism and helped change home mortgage lenders' practices. Making HMDA data mutually accessible to lending institutions and community organizations is correlated with beneficial outcomes for banks and borrowers alike [20]. However, it remains difficult to know for certain to what extent this data led to reductions in discriminatory lending practices or merely documented changes that were already underway.

2.2 Employment

United States federal law prohibits employers and employment agencies from discriminating on the basis of certain protected attributes. In this context, the collection of demographic information is a familiar part of most employers' day-to-day practice. For example, many large employers are *required* to collect demographic data about job applicants and employees to facilitate regulatory enforcement and research. And for many decades, employment selection procedures have been subject to regulatory guidelines that assume "adverse impact" can be readily quantified.

2.2.1 Background. Following sustained, nationwide demands to end racial discrimination and segregation, Congress passed sweeping protections in the Civil Rights Act in 1964. Title VII of the Act pertains specifically to employment, prohibiting employers from directly or indirectly discriminating in their employment practices and laying out expectations around data collection and reporting for enforcement purposes. The following year, President Lyndon B. Johnson signed Executive Order 11246, which prohibits federal contractors from discriminating in employment decisions, and also requires employers to take affirmative action to increase the representation of women and minorities in their workforces. The order, enforced by the Department of Labor's Office of Federal Contract Compliance (OFCCP), also outlines related requirements around the documentation of recruitment activities, including the collection of demographic information about job applicants and employees in order to facilitate the detection of discrimination at different points in the recruitment pipeline.

Title VII requires employers and other covered entities to "make and keep such records relevant to the determinations of whether unlawful employment practices have been or are being committed," as defined by the Equal Opportunity Employment Commission (EEOC), which enforces the law [1]. Since employers may be liable for employment practices that result in disparate impact on the basis of protected categories including race and gender, EEOC guidance points to Title VII as a legal basis for requiring the collection of applicant data as necessary to detect, mitigate, or defend against claims of disparate impact. The Uniform Guidelines on Employment Selection Procedures, which reflects the U.S. government's unified position on employment tests, detail how employment tests must be evaluated for unjustified adverse impact on the basis of race, sex, or ethnicity. The EEOC may allow employers to use selection procedures with disparate impact provided that the procedure has been "validated" according to these guidelines [31].

In order to support enforcement of these legal protections, monitor progress in workplace diversity, and enable employer self-assessment, the EEOC also requires private employers with 100 or more employees and contractors with more than 50 employees to collect aggregate statistics about the demographics of their workforce and report them to regulators on a yearly basis, known as EEO-1 reports.

2.2.2 Data practices. Collection of sensitive attribute data in the employment sector is highly standardized, reflecting well-defined federal reporting requirements.

For EEO-1 reports, employers must collect data on sex, a binary field (male or female), as well as race, divided into predefined categories of Black, Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native, white, or "two or more races" [16]. These categories were last updated in 2005 (after 40 years), and in 2007 the EEOC advised that employers were permitted—but not required—to collect more detailed demographic data [55, 63]. Employers must offer employees the opportunity to voluntarily self-identify in the predefined categories. If and only if an employee declines to self-identify, the employer may use "employment records or observer identification," elsewhere described as "visual surveys of the workforce" to categorize the worker to complete their reporting requirements [16].

Although not all employers are required to track sensitive attributes from job applicants, many opt to solicit this information at the time of application, and federal contractors are required to do so. Contractors may solicit demographic data from applicants at any time during the employee selection process so long as the data is solicited from all applicants. Regulators advise that "voluntary self-reporting or self-identification is still generally the preferred method for collecting data on race, ethnicity, and gender, but in situations where self-reporting is not practicable or feasible, observer information may be used to identify race, ethnicity, and gender" [62]. After making "reasonable efforts to identify applicant gender, race, and ethnicity information," contractors may record the applicant's race and gender as "unknown"—with the exception that employers may visually identify applicants "when the applicant appears in person and declines to self-identify" [61]. Notably, employers may not use these data as a part of their employment selection procedures, but may use them to evaluate outcomes and inform changes to those procedures.

2.2.3 Results and reactions. As of 2017, nearly 70 thousand employers file EEO-1 reports per year, documenting data for over 50 million employees [14]. Multiple studies have used EEO and other sources of demographic data to measure trends in occupational

segregation, finding that it has declined since the passage of Title VII [44, 69, 78]. Others use this data to more closely examine race and sex inequality in managerial positions and within specific industries, as well as gender and racial pay gaps [36, 45, 69]. Several researchers were able to determine that OFCCP monitoring and enforcement in particular likely contributed to greater representation of Black workers in skilled occupations [47]. The EEOC and OFCCP themselves commonly use EEO-1 and other mandatorily collected data to support investigations of individual and systemic employment discrimination [15, 75].

Some have pointed out that unlike other government survey instruments, the EEOC merges data on race and ethnicity, which may lead to measurement errors [69]. Others critique the allowance of observed data, but concede that because observed data relate to how workers may be perceived, these data may still have utility in understanding employment discrimination [72]. However, we identified relatively little criticsm of the overall exercise of collecting sensitive attribute data in the context of employment, perhaps because the law requiring and justifying their collection is so clear.

Here again, it is not clear that the relationship between demographic data collection and any occupational desegregation is a causal one. Without this disaggregated employment data, however, documentation of these trends would be significantly more difficult. Indeed, researchers have found that while EEO-1 data do have some constraints, they can be a particularly powerful tool to study workplace inequality and segregation, especially as compared with other data sources [69].

2.3 Health

United States federal law prohibits discrimination in the provision of various health care services. For example, those who qualify for federal health insurance programs such as Medicare or Medicaid may not be subjected to discrimination based on certain sensitive attributes. However, unlike in credit and employment, a major driving factor behind collection of sensitive attribute data in this sector has been voluntary industry efforts to address racial and ethnic disparities in health outcomes, rather than compliance with antidiscrimination laws alone.

2.3.1 Background. The passage of the Civil Rights Act in 1964 and the establishment of Medicare the following year created a need for data to confirm that patients had equal access to health care and that hospitals were not segregated. As a result, many hospitals initially collected data about sensitive attributes for compliance purposes only [68].

A shift in approach was prompted not long after by Secretary of Health and Human Services (HHS) Margaret Heckler's observation in a 1983 national health report that minority health lagged behind that of white Americans, and the subsequent formation of the Task Force on Black and Minority Health to research this gap. The 1986 publication of the Report of the Secretary's Task Force on Black and Minority Health (Heckler Report) marked the first study highlighting the significant health disparities racial minorities experienced in the U.S. [57]. Although the Heckler Report's findings drew awareness to racial inequality in healthcare provision, they did not themselves effect a shift away from compliance-based sensitive attribute data collection toward a model of using these data to reduce discrimination.

At the request of Congress in 2003, the Institute of Medicine (IOM) published a follow-up report, Unequal Treatment, affirming that unacceptable levels of racial and ethnic disparities in health outcomes persisted [60]. The IOM report concluded that without data on patients' race, ethnicity, socioeconomic status, and primary language, it would be impossible for healthcare providers to detect or address these disparities, and recommended the systematic collection and reporting of race and ethnicity data as a critical step toward eliminating them.

The IOM report jump-started health insurance and other care providers' joint, voluntary effort to collect and use data for healthcare quality improvement and disparity reduction [68]. Organizations such as the National Health Plan Collaborative (NHPC) connected health research institutes to national and regional health plans in order for the former to provide these firms with educational tools and recommendations for how to detect and mitigate discrimination [73]. While initially, many insurance providers believed collecting race and ethnicity data was illegal, legal analysis determined that collection was justified under (though not explicitly required by) Title VI of the Civil Rights Act and the Affordable Care Act, as well as several state laws. Under these statutes, health plans are prohibited from using demographic data for discriminatory purposes, including steering patients toward certain healthcare products [18, 43]. However, health plans are allowed to use these data in order to report aggregate trends and join initiatives to provide equitable services.

2.3.2 Data practices. Some health providers have found it necessary to collect data on patients' race, ethnicity, and primary spoken language (REL) to identify health care disparities [23]. However, there is substantial variability in the precise categories and level of granularity different health providers opt to use to do so. Industry-wide efforts to standardize these data are ongoing.

Physicians and hospitals often collect REL data at intake—usually by asking patients directly, though sometimes determined by intake specialist observation [28]. Health plans, on the other hand, tend to use surveys and incentive programs to collect data after people have signed up for coverage. In some cases, insurers are prohibited from asking for race/ethnicity data during the sign-up process [25, 29]. Some health providers also appear to be able to share and obtain data from federal agencies (e.g. Medicaid), though the exact mechanics of this process remain obscure.

Policymakers and practitioners recognize that in general, data that patients self-report are strongly preferred [24, 33], but in practice, providers have struggled to convince most patients to voluntarily self-report. In the interest of generating data necessary to reduce disparities, methods to estimate race and ethnicity have been widely adopted to supplement self-reported data [54]. Early inference methods involved basic geocoding and surname analysis; more advanced probabilistic techniques have since been developed to refine these estimates. These algorithms produce probabilities that individuals belong to a particular racial or ethnic group, which can then be used to assess disparities between subgroups at an

aggregate level [29, 68]. A number of health plans combine self-reported and estimated data to increase accuracy of their analysis [54].

Experts have recommended that race/ethnicity data based on indirect estimation methods should be stored separately from or be clearly marked in medical systems. Inferred data should not be placed in individuals' clinical medical records—that is, probabilistic methods should not be used to assign someone a particular race or ethnicity classification [68]—but should only be used for aggregate statistical analysis [25]. The IOM recommended that when possible, estimations should be accompanied by their respective probabilities [68]. Whether actual data management practice follows these recommendations likely varies by institution.

2.3.3 Results and reactions. While significant healthcare disparities remain, they have narrowed since the publication of the IOM report that motivated increased data collection [22, 59]. Moreover, granular data has enabled ongoing monitoring and benchmarking of health outcomes, motivated substantial scientific and policy research, and supported federal, state, local, industry, and practitioner-driven disparity reduction initiatives.

Although many health plans have internal policies on confidentiality and use of race/ethnicity data [30], low rates of participation in voluntary data collection may indicate continued lack of trust in healthcare institutions that collect these data, and fear that demographic data might be used to discriminate against patients or otherwise be misused [33]. Health plans have admitted that they sometimes hesitate to collect data for fear of being accused of discrimination [25], on top of other challenges like privacy concerns, IT limitations, and inconsistency or insensitivity in the available categories [54]. But many healthcare providers circumvent these challenges by using techniques to generate demographic data in a probabilistic manner.

Critiques of direct and indirect data collection efforts in healthcare have also emerged on the grounds that concepts of race and ethnicity are merely sociopolitical constructs [25], and therefore categorizing patients using those constructs may reinforce and calcify them. However, the broadly recognized harms of race- and ethnicity-related health disparities seem to have outweighed this critical perspective for the time being.

3 DISCUSSION

Clearly, debates about collection of sensitive attribute data for antidiscrimination purposes are not new. There are decades of precedent that can inform the machine learning fairness research community, the broader technology industry, and other stakeholders.

It is important to reiterate that our case studies do not indicate whether collection of sensitive attribute data has contributed causally to more fair and equitable outcomes. A more fulsome analysis of this question remains for future work. However, we remain convinced that measurement is often a precondition for meaningful improvements.

While the case studies above merely scratch the surface, they offer some important insights. First, they show that U.S. legal frameworks do not offer consistent, extensible guidance about when and how corporations should collect sensitive attribute data. Rather, there are divergent and sometimes contradictory approaches: Some companies are required to collect sensitive data to comply with antidiscrimination laws, while others are explicitly prohibited from doing so. Second, they show that companies' primary incentives for collecting sensitive attribute data may not—and need not—be compliance or legal requirements at all. The healthcare industry is one such example. Here, deliberate, sustained, and ongoing debates on data collection and inference practices across the industry and stakeholder communities were needed to align on an approach to combating disparities.

If awareness-based techniques remain a primary approach to bias mitigation in predictive modelling, there is a need to thoughtfully consider what efforts must be undertaken to expand collection of sensitive attribute data in a responsible manner.

3.1 Lessons regarding traditionally regulated companies

For traditionally regulated entities like banks and employers, modernization or clarification of laws and regulatory guidance may be needed to encourage the collection of sensitive attribute data for new antidiscrimination efforts. Because these companies can be liable for discriminatory outcomes, they are unlikely to voluntarily collect or analyze sensitive attribute data that could introduce new vectors for liability. Thus, they might resist legal reforms that make it easier to collect sensitive attribute data.

Looking ahead, policymakers, researchers, and and civil society will need to work together to assess what kinds of sensitive attribute data are needed to protect people against discrimination *and* create the policy conditions for that collection to occur. These stakeholders will need to consider what data ought to be collected and in what form, and the appropriate scope of "safe harbor" provisions to incentivize thorough and transparent study. These are not clear or settled questions, even with decades of practice under longstanding civil rights laws.

3.2 Lessons regarding less regulated companies

Many technology platform companies, including those using models to mediate access to important life opportunities, are not squarely covered by civil rights laws. These companies often operate as internet intermediaries, and thus enjoy some special legal protections from liability arising from content posted by third party users [26]. Nonetheless, many are grappling with how to prevent bias. For example:

- Airbnb recently assembled "a permanent team of engineers, data scientists, researchers, and designers whose sole purpose is to advance belonging and inclusion and to root out bias" [52]. The announcement came on the heels of reports of discrimination against African Americans seeking housing opportunities on its platform. The company has not yet publicly discussed the details of this work, or whether it collects or infers sensitive attribute data in its efforts to combat discrimination. However, it is difficult to imagine an approach that would avoid these questions.
- Facebook, in delivering advertisements on its platform, introduces demographic skews along gender and race lines [5]. This practice is currently being challenged in court by the

United States Department of Housing and Urban Development (HUD) [58]. Furthermore, as part of a legally enforceable settlement with civil rights organizations, Facebook recently committed to studying the potential for unintended biases in algorithmic modeling [6]. However, this research will likely be impossible without collecting or inferring sensitive attributes of the company's users. It is not yet clear how Facebook will approach this issue.

- LinkedIn, in an effort to promote equity in hiring, recently updated its recruiter tools to balance the gender distribution in candidate search results, rather than sorting candidates purely by "relevance" [12]. With this update, if the pool of potential candidates who fit an employer's search parameters reflects a certain proportion of women, LinkedIn will re-rank candidates so that every page of search results reflects that proportion. The company also plans to offer employers reports that track the gender breakdown of their candidates across several stages of the recruitment process, as well as comparisons to the gender makeup of peer companies. These features rely on inferring gender data about jobseekers on the platform, which the company was already doing for advertising purposes.

It's not surprising that each of the above examples was motivated by some combination of public pressure or litigation. Technology companies are unsure about what kinds of sensitive attribute collection are appropriate. As a result, the path of least resistance is to simply not to collect or infer data that may create controversy or highlight disparities that may be difficult to address. This is especially true given that perceived violations of privacy are likely to garner intensive media coverage, or where applicable, increased attention from regulators. It will likely fall to a wide range of stakeholders, including advocates, researchers, and policymakers, to ensure that sensitive attribute data is collected and used under appropriate circumstances.

3.3 The need for multidisciplinary collaboration

The implementation of awareness-based antidiscrimination approaches cannot, and should not, move forward without robust involvement of public interest, technical, and regulatory stakeholders. Even amid clear and compelling risks of discrimination or unjust demographic disparities, it can be difficult for policymakers to recommend the collection of sensitive attribute data. There is no evidence this issue will become easier in the future, despite the rapid adoption of machine learning models involved in important life decisions for which these data may be critical to prevent harm.

Privacy laws can sometimes sit in tension with antidiscrimination goals, and might prevent well-meaning actors from collecting data that are necessary to detect and remediate bias in machine learning-based models. Privacy advocates will need to ensure that new legal requirements around data minimization and restrictions on the processing of sensitive data do not deter or impede companies from good faith self-testing and bias remediation. At least one recent U.S. legislative proposal provides an explicit exception for such testing [2], reinforcing the need for more detailed implementation guidance. Meanwhile, European laws and norms diverge significantly from the U.S. approach, prioritizing privacy heavily over awareness-based antidiscrimination approaches [79]. Private entities may need to navigate conflicting laws, guidance, and public expectations across social and geopolitical contexts.

Finally, there is no shortage of critical questions that still need to be answered:

- *When should sensitive attribute data be collected?* Given the practices described above, non-industry actors should consider under what conditions, if any, they would trust certain private actors with sensitive attribute data that are needed for antidiscrimination efforts. It's obvious that data collection would be justified in some contexts, but the risks may outweigh potential benefits in others. It's far less obvious (and beyond the scope of this paper to suggest) where those lines should be drawn. These norms are especially unsettled for technology companies, who have not had the same historical obligations as traditionally regulated entities, and suffer from significant trust deficits around their data practices.
- *How should sensitive attribute data be created?* Sensitive attribute data can be collected directly from subjects or inferred from non-sensitive data. However, inference presents challenges around consent, forced classification, and error. Stakeholders must work together to determine under what conditions inference is acceptable, appropriate inference methodologies, and how to treat inferred data responsibly. The cases considered in this paper offer instructive approaches, including retaining probabilistic values and uncertainty in inferred data, clearly marking when data are observed or inferred, and storing inferred data separately from data collected with permission. Other approaches might include enforceable commitments to use these data only for detecting and mitigating discrimination.
- *How should sensitive attribute data be treated and secured?* Ideally, sensitive data would be stored separately from other data and used only for limited purposes, but such technical safeguards may be difficult to guarantee. New privacy-protective techniques to access sensitive attribute data, including secure multi-party computation tools like private set intersection and homomorphic encryption, may allow companies to securely sequester these sensitive data from general purpose user data, or even enable trusted third parties to collect, infer, or hold sensitive data while making their insights available to the private entities whose products implicate people's rights [76]. However, these techniques are still nascent and have yet to be broadly deployed for the purpose of bias testing or mitigation.

4 CONCLUSION

Policy debates about the collection and use of sensitive attribute data will decide the fate of awareness-based bias mitigation techniques. There is an urgent need for machine learning scholars to drive these conversations forward, along with other stakeholders, so policy and technical approaches can be developed in accordance with each other. The ability to detect and address bias in algorithms—and the durability of foundational civil rights protections—may hang in the balance.

REFERENCES

[1] [n.d.]. *Title VII, SEC. 2000e-8. [Section 709].* Retrieved August 21, 2019 from https://www.eeoc.gov/laws/statutes/titlevii.cfm

[2] 2019. Consumer Online Privacy Rights Act (COPRA). https://www.cantwell.senate.gov/imo/media/doc/COPRA%20Bill%20Text.pdf

[3] J. Khadijah Abdurahman. 2019. *FAT* Be Wilin': A Response to Racial Categories of Machine Learning by Sebastian Benthall and Bruce Haynes.* Retrieved August 21, 2019 from https://medium.com/@blacksirenradio/fat-be-wilin-deb56bf92539

[4] Grace Abuhamad. 2019. *The Fallacy of Equating "Blindness" with Fairness: Ensuring Trust in Machine Learning Applications to Consumer Credit.* Master's thesis. MIT Institute for Data, Systems, and Society. https://dspace.mit.edu/handle/1721.1/122094

[5] Muhammad Ali, Aleksandra Korolova, Piotr Sapiezynski, Alan Mislove, Miranda Bogen, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. (2019), 1–15. https://arxiv.org/abs/1904.02095

[6] National Fair Housing Alliance. 2019. *Exhibit A: Programmatic Relief.* Retrieved August 21, 2019 from https://nationalfairhousing.org/wp-content/uploads/2019/03/FINAL-Exhibit-A-3-18.pdf

[7] AnnaMaria Andriotis and Rachel Louise Ensign. 2015. *U.S. Government Uses Race Test for USD80 Million in Payments.* Retrieved August 21, 2019 from https://www.wsj.com/articles/u-s-uses-race-test-to-decide-who-to-pay-in-ally-auto-loan-pact-1446111002

[8] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671–732. http://dx.doi.org/10.15779/Z38BG31

[9] Google AI Blog. 2018. *The What-If Tool: Code-free probing of machine learning models.* Retrieved August 21, 2019 from https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 81. 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[11] Consumer Financial Protection Bureau. 2014. *Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment.* Retrieved August 21, 2019 from https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[12] Rosalie Chan. 2019. *LinkedIn is using AI to make recruiting diverse candidates a no-brainer.* Retrieved August 21, 2019 from https://www.businessinsider.com/linkedin-new-ai-feature-increase-diversity-hiring-2018-10

[13] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 339–348. https://doi.org/10.1145/3287560.3287594

[14] U.S. Equal Employment Opportunity Commission. [n.d.]. *Job Patterns For Minorities And Women In Private Industry (EEO-1).* Retrieved August 21, 2019 from https://www.eeoc.gov/eeoc/statistics/employment/jobpat-eeo1/

[15] U.S. Equal Employment Opportunity Commission. 2016. *Advancing Opportunity - A Review of the Systemic Program of the U.S. Equal Employment Opportunity Commission: Recruitment and Hiring.* Retrieved August 21, 2019 from https://www.eeoc.gov/eeoc/systemic/review/index.cfm#IIIA

[16] U.S. Equal Employment Opportunity Commission. 2018. *Equal Employment Opportunity Standard Form 100, Employer Information Report EEO-1 Instruction Booklet.* Retrieved August 21, 2019 from https://www.eeoc.gov/employers/eeo1survey/2007instructions.cfm

[17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. 797–806. https://doi.org/10.1145/3097983.3098095

[18] National Research Council. 2004. *Eliminating Health Disparities: Measurement and Data Needs.* Retrieved December 4, 2019 from https://www.nap.edu/catalog/10979/eliminating-health-disparities-measurement-and-data-needs

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*. 214–226. https://doi.org/10.1145/2090236.2090255

[20] National Consumer Law Center et al. 2015. *Group Letter to Consumer Financial Protection Bureau Regarding Public Disclosure of New HMDA Data Points.* Retrieved August 21, 2019 from https://www.nclc.org/images/pdf/foreclosure_mortgage/predatory_mortgage_lending/letter-re-hmda-benefits-and-privacy.pdf

[21] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. https://doi.org/10.1145/2783258.2783311

[22] National Center for Health Statistics. 2016. Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities. (May 2016), iii–449. https://www.ncbi.nlm.nih.gov/books/NBK367640/pdf/Bookshelf_NBK367640.pdf

[23] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata1a.html

[24] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - 1. Introduction.* Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata1.html

[25] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - Summary.* Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldatasum.html

[26] 47 U.S. Code Section 230. Protection for private blocking and screening of offensive material. [n.d.]. *Section 230. Protection for private blocking and screening of offensive material.* Retrieved August 21, 2019 from https://www.law.cornell.edu/uscode/text/47/230

[27] James Foulds and Shimei Pan. 2018. An Intersectional Definition of Fairness. (2018), 1–16. https://arxiv.org/abs/1807.08362

[28] Robert Wood Johnson Foundation. 2011. *Moving Toward Racial and Ethnic Equity in Health Care.* Retrieved August 21, 2019 from https://www.rwjf.org/en/library/research/2011/03/moving-toward-racial-and-ethnic-equity-in-health-care.html

[29] Allen Fremont, Joel S. Weissman, Emily Hoch, and Marc N. Elliott. [n.d.]. *When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities.* Retrieved August 21, 2019 from https://www.rand.org/pubs/periodicals/health-quarterly/issues/v6/n1/16.html

[30] Julie Gazmararian, Rita Carreon, Nicole Olson, and Barbara Lardy. 2012. Exploring Health Plan Perspectives in Collecting and Using Data on Race, Ethnicity, and Language. *The American Journal of Managed Care* 18, 7 (July 2012), e254–e261. https://www.ncbi.nlm.nih.gov/pubmed/22823554

[31] Biddle Consulting Group. [n.d.]. *Uniform Guidelines on Employee Selection Procedures, Section 4: Information on impact.* Retrieved August 21, 2019 from http://uniformguidelines.com/uniformguidelines.html#15

[32] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS '16)*. 3323–3331. https://dl.acm.org/citation.cfm?id=3157469

[33] Romana Hasnain-Wynia and David W Baker. 2006. Obtaining Data on Patient Race, Ethnicity, and Primary Language in Health Care Organizations: Current Challenges and Proposed Solutions. *Health Services Research* 41 (Aug. 2006), 1501–1518. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1797091/

[34] Anna Lauren Hoffman. 2018. Data, technology, and gender: Thinking about (and from) trans lives (as Incoll). In *Spaces for the Future: A Companion to Philosophy of Technology*. Routledge, New York, NY, USA.

[35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daume III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–16. https://doi.org/10.1145/3290605.3300830

[36] Matt L. Huffman, Philip N. Cohen, and Jessica Pearlman. 2010. Engendering Change: Organizational Dynamics and Workplace Gender Desegregation, 1975âĂŞ2005. *Administrative Science Quarterly* 55, 2 (June 2010), 255–277. https://doi.org/10.2189/asqu.2010.55.2.255

[37] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 49–58. https://doi.org/10.1145/3287560.3287600

[38] IBM. 2019. *AI Fairness 360 toolkit.* Retrieved August 21, 2019 from https://github.com/IBM/AIF360

[39] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[40] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. 1–5. https://arxiv.org/abs/1711.05144

[41] Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. 2630–2639. https://arxiv.org/abs/1806.03281

[42] Joseph M. Kolar and Jonathan D. Jerison. 2006. *The Home Mortgage Disclosure Act: Its History, Evolution, and Limitations.* Retrieved August 21, 2019 from https://buckleyfirm.com/uploads/36/doc/HistoryofHMDAapr06.pdf

[43] Sarah Kornblet, Joy Prittsa, Melissa Goldstein, Tom Perez, and Sara Rosenbaum. [n.d.]. Patient Race And Ethnicity Data And Quality Reporting: A Legal. ([n. d.]).

[44] Fidan Ana Kurtulus. 2012. Affirmative Action and the Occupational Advancement of Minorities and Women During 1973-2003. *Industrial Relations* (April 2012),

213–246. https://doi.org/10.1111/j.1468-232X.2012.00675.x
[45] Fidan Ana Kurtulus and Donald Tomaskovic-Devey. 2012. Do Female Top Managers Help Women to Advance? A Panel Study Using EEO-1 Records. *Annals of American Academy of Political and Social Science* 639 (Jan. 2012), 173–197. https://ssrn.com/abstract=1999538
[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm.* Retrieved August 21, 2019 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[47] Jonathan S. Leonard. 1984. Employment and Occupational Advance Under Affirmative Action. *The Review of Economics and Statistics* 66, 3 (Aug. 1984), 377–385. https://www.jstor.org/stable/1924993
[48] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 349–358. https://doi.org/10.1145/3287560.3287564
[49] Richard D. Marsico. 1999-2000. Shedding Some Light on Lending: The Effect of Expanded Disclosure Laws on Home Mortgage Marketing, Lending and Discrimination in the New York Metropolitan Area. *Fordham Urban Law Journal* 481 (1999-2000), 484. https://pdfs.semanticscholar.org/a009/82327779e40af15816c6cb87d027af40e593.pdf
[50] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of Machine Learning Research*, Vol. 81. 1–12. https://arxiv.org/pdf/1705.09055.pdf
[51] Emanuel D. Moss. 2019. Translation Tutorial: Toward a Theory of Race for Fairness in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 1–2. https://drive.google.com/file/d/1jKSbnAg7NPotjhvNb8o4YXyjcjF8RTJW/view
[52] Laura Murphy. 2016. *Airbnb's Work to Fight Discrimination and Build Inclusion: A Report Submitted to Airbnb.* Retrieved August 21, 2019 from https://blog.atairbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf
[53] Arvind Narayanan. 2008. 21 Fairness Definitions and Their Politics. Tutorial presented at the Conf. on Fairness, Accountability, and Transparency.
[54] David R. Nerenz, Rita Carreon, and German Veselovskiy. 2013. Race, ethnicity, and language data collection by health plans: findings from 2010 AHIPF-RWJF survey. *Journal of Health Care for the Poor and Underserved* 24, 4 (Nov. 2013), 1769–1783. https://muse.jhu.edu/article/524354/pdf
[55] Jeffrey A. Norris. 2007. *EEOC Revises Its EEO-1.*
[56] Board of Governors of the Federal Reserve System. 2007. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit.* Retrieved August 21, 2019 from https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf
[57] U.S. Department of Health and Human Services. [n.d.]. *Report of the Secretary's Task Force on Black and Minority Health.* Retrieved August 21, 2019 from https://minorityhealth.hhs.gov/heckler30/
[58] Department of Housing and Urban Development. 2018. *Department of Housing and Urban Development Charge of Discrimination Against Facebook, Inc.* Retrieved August 21, 2019 from https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf
[59] Institute of Medicine. 2012. *How far have we come in reducing health disparities?: Progress since 2000: Workshop summary.* The National Academies Press, Washington, DC.
[60] Institute of Medicine Committee on Understanding, Eliminating Racial, and Ethnic Disparities in Health Care. [n.d.]. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Retrieved August 21, 2019 from https://www.ncbi.nlm.nih.gov/pubmed/25032386
[61] Federal Contract Compliance Programs Office. 2005. *Obligation To Solicit Race and Gender Data for Agency Enforcement Purposes.* Retrieved August 21, 2019 from https://www.federalregister.gov/documents/2005/10/07/05-20176/obligation-to-solicit-race-and-gender-data-for-agency-enforcement-purposes
[62] U.S. Department of Labor Office of Federal Contract Compliance Programs. [n.d.]. *Internet Applicant Recordkeeping Rule.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/regs/compliance/faqs/iappfaqs.htm#Q7GI
[63] U.S. Department of Labor Office of Federal Contract Compliance Programs. 2008. *OFCCP Directive No. 283.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/regs/compliance/directives/dir283.htm
[64] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 1–10. https://doi.org/10.1145/3287560.3287567
[65] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. 560–568. https://doi.org/10.1145/1401890.1401959
[66] Corinne Purtill. 2015. *The US is using a race-predicting algorithm to identify victims of car loan discrimination.* Retrieved August 21, 2019 from https://qz.com/537816/in-an-auto-loan-discrimination-case-a-race-predicting-algorithm-is-the-us-governments-best-shot-at-paying-victims-back/
[67] Pymetrics. 2019. *audit-AI (Python library).* Retrieved August 21, 2019 from https://github.com/pymetrics/audit-ai
[68] Agency For Healthcare Research and Quality. [n.d.]. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - 3. Defining Categorization Needs for Race and Ethnicity Data.* Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata3.html
[69] Corre L. Robinson, Tiffany Taylor, Donald Tomaskovic-Devey, Catherine Zimmer, and Jr. Matthew W. Irvin. 2005. Studying Race or Ethnic and Sex Segregation at the Establishment Level: Methodological Issues and Substantive Opportunities Using EEO-1 Reports. *Work and Occupations* 32, 1 (Feb. 2005), 5–38. https://doi.org/10.1177/0730888404272008
[70] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 59–68. https://doi.org/10.1145/3287560.3287598
[71] Lynn Severson-Meyer. 2003. *Recent ECOA amendments allow lenders an exception for self-testing.* Retrieved August 21, 2019 from https://www.minneapolisfed.org/publications/community-dividend/recent-ecoa-amendments-allow-lenders-an-exception-for-selftesting
[72] Patrick Simon. 2005. The measurement of racial discrimination: the policy use of statistics. *International Social Science Journal* 57, 183 (March 2005), 9–25. https://doi.org/10.1111/j.0020-8701.2005.00528.x
[73] Erin Fries Taylor and Marsha Gold. [n.d.]. *The National Health Plan Collaborative: Overview of Its Origins, Accomplishments, and Lessons Learned.* Retrieved August 21, 2019 from https://www.ahrq.gov/sites/default/files/publications/files/nhpceval.pdf
[74] Winnie Taylor. 2011-12. Proving Racial Discrimination and Monitoring Fair Lending Compliance: The Missing Data Problem in Nonmortgage Credit. *Review of Banking and Financial Law* 31 (2011-12), 218. https://heinonline.org/HOL/P?h=hein.journals/annrbfl31&i=208
[75] Office of Federal Contract Compliance Programs U.S. Department of Labor. [n.d.]. *OFCCP By The Numbers.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/BTN/index.html
[76] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society* (July 2017), 1–17. https://doi.org/10.1177/2053951717743530
[77] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. 1–14. https://doi.org/10.1145/3173574.3174014
[78] Kim A. Weeden, Mary Newhart, and Dafna Gelbgiser. 2018. State of the Union 2018: Occupational Segregation. *The Poverty and Inequality Report (Stanford Center on Poverty and Inequality* (2018), 30–33. https://inequality.stanford.edu/sites/default/files/Pathways_SOTU_2018_occupational-segregation.pdf
[79] Indre Zliobaite and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (June 2016), 183–201. https://link.springer.com/article/10.1007/s10506-016-9182-5

Fair Isaac Corporation
200 Smith Ranch Road
San Rafael, CA 94903 USA
T 415 446 6000
F 415 492 9381
www.fico.com

July 27, 2020

NAIC Casualty Actuarial and Statistical Task Force
Attn: Kris DeFrain, FCAS, MAAA, CPCU

Via Email – kdefrain@naic.org

Re: Comments on the June 12, 2020 Exposed CAS Task Force Draft White Paper –
Best Practices – Regulatory Review of Predictive Models

Dear Ms. DeFrain:

Fair Isaac Corporation (FICO) appreciates the opportunity to submit these brief comments on the most recently released draft white paper, *Best Practices – Regulatory Review of Predictive Models*.

FICO fully understands and respects the value of regulatory scrutiny as well as the need for regulatory flexibility to help ensure that consumers continue to enjoy greater access to more affordable insurance through the industry's use of credit-based insurance scores. We believe it is important to preserve this consumer benefit.

FICO remains concerned that the predictive model review provisions proposed in the white paper, if adopted and implemented by state departments of insurance, will reduce the effectiveness of the time-tested and proven regulatory processes that exists today with respect to credit-based insurance scores. FICO believes that implementation of the proposed provisions is likely to severely strain the strong competitive nature of the auto and home insurance industry, which leads to lower insurance costs for consumers. Greater competition drives to greater access of insurance and at lower costs for consumers.

More importantly, however, the majority of consumers who have benefitted—via lower premium payments—from the industry's use of credit-based insurance scores for the past two decades will see their premiums increase as this key risk segmentation tool is restrained or restricted in those states that adopt the proposed provisions. Properly utilized and proven risk segmentation tools allow insurers to identify those risks requiring more or less premium based on the likelihood of future claim activity.

For nearly two decades, in support of successful rate filings throughout the nation by our FICO® Insurance Score clients, FICO has provided model documentation to all requesting departments of insurance that are able to provide appropriate confidentiality protections for FICO's proprietary information. FICO's submissions generally include specific consumer credit characteristics, attributes and weights for the filed model, reason code/factor definitions, and a general discussion of our model development process.

FICO will continue to provide that filing support for our clients' use of FICO Insurance Scores by answering all appropriate regulatory questions to the best of our ability and by offering as much insight into FICO's proprietary modeling analytics and technologies as possible, while still protecting our intellectual property.

Thank you for allowing FICO to once again comment on this important issue.

Sincerely,

Lamont D. Boyd, CPCU, AIM
Insurance Industry Director, Scores
FICO Decisions
LamontBoyd@FICO.com
602-317-6143 (mobile)
FICO® Insurance Scores Consumer website - https://insurancescores.fico.com/

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\FICO Comments

Milliman

650 California Street, 21st Floor
San Francisco, California 94108

Tel +1 415 403-1333
Fax +1 415 403-1334

www.milliman.com

July 27, 2020
Ms. Kris DeFrain, FCAS, MAAA, CPCU
Director, Research and Actuarial
NAIC Central Office

Re: CASTF Predictive Model White Paper – Proposed Edits

Dear Ms. DeFrain:

In support of the finalization of the *Regulatory Review of Predictive Models White Paper*, we respectfully submit the following suggested edits.

COMMENTS

Our comments are summarized in the following table:

Page	Section	Exposure Draft Text	Comment
11	X. Other Considerations	Provide guidance for regulators on the value and/or concerns of data mining, including how data mining may assist in the model building process, how data mining may conflict with standard scientific principles, how data mining may increase "false positives" during the model building process, and how data mining may result in less accurate models or models that are unfairly discriminatory.	Previous exposure draft comments have mentioned the potential issues of brief mentions of other topics that suggest concerns without sufficient discussion. "Data mining" is no longer defined within the paper, and we believe some readers may associate "data mining" with the general application of machine learning and statistical models. Done with appropriate validation, this use of "data mining" is not inherently problematic. We suggest this item be reworded to refer to "data dredging" or "the use of predictive modeling methods without sufficient validation."
14	Appendix B, paragraph 2	Documentation should be sufficiently detailed and complete to enable a qualified third party to form a sound judgment on the suitability of the model for the intended purpose.	We suggest adding guidance based on ASOP 56 - *Modeling*, clarifying that "the degree of such documentation may vary with the complexity and purpose of the model."
16	A.1.a	If the data is taken from an outside source, find out what steps were taken to verify the data was accurate, complete and unbiased in terms of relevant and representative time frame, representative of potential exposures and lacking in obvious correlation to protected classes.	We suggest an example to illustrate what would be considered "obvious correlation."
18	A.2.e	Such redundancy may also occur with the inclusion of fluvial or pluvial flood losses when using a flood model, inclusion of freeze losses when using a winter storm	It is not clear how including demand surge on catastrophe model output would create overlap/redundancy with historical claims data. We suggest a clarification, or the

		model or including demand surge caused by any catastrophic event.	removal the reference to demand surge.
18	A.2.e	Note that, the rating plan or indications underlying the rating plan, may provide special treatment of large losses and non-modeled large loss events. If such treatments exist, the company should provide an explanation how they were handled. These treatments need to be identified and the company/regulator needs to determine whether model data needs to be adjusted. For example, should large BI losses, in the case of personal automobile insurance, be capped or excluded, or should large non-catastrophe wind/hail claims in home insurance be excluded from the model's training, test and validation data?	This comment appears in sections A.2.e and A.3.a. We suggest removing it from A.2.e as it is not related to the issue of redundancy between modeled losses and historical claims.
22	B.1.g	The modeler should comment if any form of data mining to identify selected variables was performed and explain how the modeler addressed "false positives" which often arise from data mining techniques.	The term "data mining" is used without definition. We suggest replacing with: "The modeler should comment on the use of automated feature selection algorithms to choose predictor variables, and explain how potential overfitting which can arise from these techniques was addressed."
22	B.1.h	In conjunction with variable selection, obtain a narrative on how the company determine the granularity of the rating variables during model development.	Replace "determine" with "determined."
25	B.4.b	For all variables (discrete or continuous), review the appropriate parameter values, confidence intervals, chi-square tests, p-values and any other relevant and material tests. Determine if model development data, validation data, test data or other data was used for these tests.	The "and" suggests that the best practice is to review not only multiple but every possible test of significance. We believe this unnecessary and impractical, and suggest that the information element should be reworded to "review the appropriate parameter values and relevant tests of significance, such as confidence intervals, chi-square tests, p-values, or F tests."
26	B.4.d	For overall discrete variables, review type 3 chi-square tests, p-values, F tests and any other relevant and material test. Determine if model development data, validation data, test data or other data was used for these tests.	Please clarify how this information element differs from B.4.b.
26	B.4.f	For continuous variables, provide confidence intervals, chi-square tests, p-values and any other relevant and material test. Determine if model development data, validation data, test data or other data was used for these tests.	Please clarify how this information element differs from B.4.b.
27	B.4.g	Obtain a description how the model was tested for stability over time.	Examples of how a modeler would perform such tests would help clarify what is meant by "stability over time." Is this

			referring to potential distortions in the historical data (which could be identified by examining lift charts by year, for example) or the ability of the model to predict on a time period independent of the training data, or both?
28	B.5.b	It is expected that there should be improvement in the Gini coefficient.	This element seems to suggest that the Gini coefficient is a "gold standard" for model selection. We suggest a clarification that the choice of a final model may be based on other technical measures of model performance, as well as business considerations.
31	C.4.c	A more granular rating plan implies that the insurer had to extrapolate certain rating treatments, especially at the tails of a distribution of attributes, in a manner not specified by the model indications.	A more granular rating plan may also arise if the insurer interpolates factors for a continuous variable that was binned into a discrete variable for modeling purposes.
32	C.7.a	Obtain a listing of the top five rating variables that contribute the most to large swings in premium, both as increases and decreases.	The comments do not provide much guidance on how an insurer or regulator would determine these variables in practice. An example would be helpful.
32	C.7.c	For the proposed filing, obtain the impacts on expiring policies and describe the process used by management, if any, to mitigate those impacts.	Item C.7.c refers to "expiring policies" and item C.7.d refer to "renewal business (created by rerating the current book of business)". Are these referring to the same thing (the inforce book of business)? If so it would be clearer to use consistent terminology. If not, please provide a comment that clarifies the difference between the two.
33	C.7.d	Obtain a rate disruption/dislocation analysis, demonstrating the distribution of percentage and/or dollar impacts on renewal business (created by rerating the current book of business), and sufficient information to explain the disruptions to individual consumers. See Appendix C for an example of a disruption analysis.	It is not clear what it required for "sufficient information to explain the disruptions to individual consumers." Since this is a Level 2 item, examples would be helpful for insurers that seek speed to market. Does the example analysis in the Appendix provide "sufficient information to explain the disruptions to individual consumers," or is more required? Also "Appendix C" should be changed to "Appendix D."
33	C.7.e	See Appendix C for an example of an exposure distribution.	Revise "Appendix C" to "Appendix D."

CLOSING

We appreciate the opportunity to comment on the exposure draft. Please feel free to reach out to me at (415) 394-3726 or peggy.brinkmann@milliman.com with any questions about the comments.

Sincerely,

Peggy Brinkmann, FCAS, MAAA, CSPA
Principal & Consulting Actuary

Nancy P. Watkins, FCAS, MAAA
Principal & Consulting Actuary

Cody Webb, FCAS, MAAA
Principal & Consulting Actuary

Greg Dietzen, FCAS, MAAA
Consulting Actuary

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\\Milliman Comments

NAMIC®
NATIONAL ASSOCIATION OF MUTUAL INSURANCE COMPANIES

317.875.5250 | [F] 317.879.8408
3601 Vincennes Road, Indianapolis, Indiana 46268

202.628.1558 | [F] 202.628.1601
20 F Street N.W., Suite 510 | Washington, D.C. 20001

July 27, 2020

NAIC Casualty Actuarial and Statistical (C) Task Force
c/o Kris DeFrain - kdefrain@naic.org
1100 Walnut Street, Suite 1500
Kansas City, MO 64106-2197

Re: NAMIC Comments on CASTF's Predictive Model White Paper – 6/12/2020 Exposure

Dear Chair and Members of the Task Force,

Please find included herein the following comments of the National Association of Mutual Insurance Companies (hereinafter "NAMIC")[1] regarding the June 12, 2020 exposure draft of the Casualty Actuarial and Statistical Task Force (CASTF) Regulatory Review of Predictive Models White Paper. NAMIC wishes to thank the task force for the ability to provide additional comments on the white paper as amended. As mentioned in previous comments, dealing with the concept of predictive modeling is an extremely important undertaking as it impacts industry and consumers. NAMIC is cognizant of the need to assure that there are appropriate parameters around predictive model usage to ensure appropriate implementation.

Throughout this more than year and one-half process, there have been many revisions and detailed analysis concerning the paper contents and comments submitted. NAMIC wishes to thank CASTF for its thorough and granular review of the submitted comments and its substantive review and response to each aspect of the submissions. The work of CASTF has been exemplary from a transparency and responsiveness viewpoint.

Woven into this discussion is also the need to recognize the current national discussion of race and discrimination that has rightfully permeated public discourse. Additionally, with the ongoing COVID-19 pandemic causing widespread loss to the world, CASTF has no doubt been mindful of these events and wants to center discussion on being responsive to consumer and regulatory concerns. NAMIC is firmly against intentional and unfair discrimination against any protected class and adamantly supports any efforts to eliminate such conduct. We believe that current regulatory authority is more than ample in assisting regulators in performing their required duties under the law but understand and appreciate the need to explore further.

[1] NAMIC membership includes more than 1,400 member companies. The association supports regional and local mutual insurance companies on main streets across America and many of the country's largest national insurers. NAMIC member companies write $278 billion in annual premiums. Our members account for 58 percent of homeowners, 44 percent of automobile, and 30 percent of the business insurance markets. Through our advocacy programs we promote public policy solutions that benefit NAMIC member companies and the policyholders they serve and foster greater understanding and recognition of the unique alignment of interests between management and policyholders of mutual companies.

NAMIC also firmly supports risk-based pricing for insurance products that use facially neutral rating factors to objectively base pricing on a person's correlated risk of future loss. This has been the insurance dynamic for many years. These discussions should and will continue and we are confident that common ground can be found in these areas with further elucidation.

The task force is charged with producing a white paper that helps provide guidelines and parameters to rate and form filings concerning predictive model usage for homeowner and auto personal lines insurance. We understand that this work is nearing its end and are mindful of the Chair's admonition not to continue to re-visit prior comments. Therefore we would respectfully like to limit our comments to concerns with the new exposure draft and only very briefly note that some of our concerns from the beginning continue through the current draft so as to avoid any misinterpretation that we no longer believe they are issues that should be addressed.

First of all, we appreciate that there have been many positive changes to the paper such as dividing the informational elements into categories of need, removing extraneous or redundant elements, and attempting to clarify or remove some potentially problematic terminology. CASTF has also endeavored to remove issues from the paper that are beyond its scope and list topics for future discourse. These are positive and again we want to thank CASTF for the effort in this regard.

However, we would recommend additional edits to make the paper the best product for review and usage by the states in their respective roles and duties. NAMIC strongly supports clear and fair regulatory guidance that protects consumers and provides a level playing field for all participants.

With those principles in mind, we are concerned that the use of "rational explanation" does not clear up the issue of correlation versus causation. In fact, in the explanatory notes in information element B.3.d, it states "[t]he explanation should go beyond demonstrating correlation." It goes on further to state that "[i]f no rational explanation can be provided; greater scrutiny may be appropriate." This interjects a highly subjective standard and moves away from traditional actuarial justification based upon correlation. The paper goes on to state in C.2.a, that a narrative should include "a rational relationship to cost." Finally, in the "Rational Explanation" definition in the Glossary of Terms the paper states –" A 'rational explanation' refers to a plausible narrative connecting the variable and/or treatment in question with real-world circumstances or behaviors that contribute to the risk of insurance loss in a manner that is readily understandable to a consumer or other educated layperson. A "rational explanation" does not require strict proof of causality but should establish a sufficient degree of confidence that the variable and/or treatment selected are not obscure, irrelevant, or arbitrary."

The use of rational explanation is still a move to a heightened standard of actuarial justification from existing law. Rates must be adequate, not excessive, or unfairly discriminatory according to state laws across the country. Actuarial soundness is dependent upon correlation to loss. We appreciate that the paper stops short of recommending a causality standard, an impossible standard where there must be absolute linkage between a variable and loss (creating an expectation that insurers

accurately and with absolute precision predict the future). A rational explanation, utilizing the provided definition, does not protect against subjective interpretation and ambiguity and therefore does not meet the goal of providing clear regulatory guidance. Therefore, we respectfully would submit that the statements concerning rational relationship and the movement beyond correlation be removed, or at least moved to the list of issues for further consideration.

Mindful of the Chair's admonition, we will only briefly touch upon concerns that we have previously raised but remain. In this regard, NAMIC would highlight among others:

- The prescriptive nature of the document is more akin to a model law or regulation and therefore should proceed in that fashion as opposed to a white paper;
- The remaining information elements (especially categories 1 and 2) still may cause a significant compliance burden, potentially slowing speed to market and ultimately stifling innovation;
- Broad terminology such as "improve" the rating plan or make the rating plan "fairer" can mean many different things to different people. Such words are not necessary to carry out the intent of the paper and may be creating a new standard;
- The sharing of proprietary models could cause irreparable harm to insurers' research and development of suitable products (we appreciate the added language on this issue, but it is still unclear how this filing information will be utilized and potentially shared inside NAIC and with the states); and
- The need for the inclusion of robust confidentiality protections for proprietary information by regulators.[2]

Finally, and in summary, we mention these items not to be repetitive or unmindful of the concerns of CASTF, but to emphasize that the eventual output of NAIC and CASTF may be utilized or interpreted improperly which we are certain is not the intent. However, the paper cannot be summarily characterized as merely guidance or best practices that states can ignore. We believe that NAIC documents may be utilized in litigation or in administrative reviews to inappropriately apply standards that are not found in law. NAIC's actions carry a great deal of weight in the public domain which is why NAMIC remains so focused on tethering NAIC work products – including aspirational products – to existing authority lest confusion develop.

NAMIC wants to thank the task force for its diligent and thorough process, the ability to respond to various drafts, and looks forward to providing continued input and finding common ground in this endeavor. We believe with a few, but significant surgical corrections the document can provide its intended goal without sacrificing the clarity and direction it intends to provide.

[2] NAMIC collaborated on a white paper concerning the CASTF white paper entitled, *The State Rating Statutes and Constitutional Policymaking: Causation and Disparate Impact Standards in NAIC's Draft White Paper*, Shapo, 2020, that discusses these and other issues, although there have been updates to the CASTF paper since publication. *See* https://www.namic.org/pdf/publicpolicy/200309_castf_issueanalysis_final.pdf.

Sincerely,

Andrew Pauley, CPCU
Government Affairs Counsel
National Association of Mutual Insurance Companies (NAMIC)

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\NAMIC Comments

PINNACLE
ACTUARIAL RESOURCES, INC.

2410 Camino Ramon, Suite 346
San Ramon, CA 94583
415.692.0938
pinnacleactuaries.com

Laura A. Maxwell, FCAS, MAAA, CSPA
Senior Consulting Actuary
LMaxwell@PinnacleActuaries.com

July 27, 2020

Kris DeFrain, FCAS, MAAA, CPCU
Director, Research and Actuarial Services
National Association of Insurance Commissioners

Sent via email: kdefrain@naic.org

Re: National Association of Insurance Commissioner's Casualty Actuarial and Statistical Task Force Regulatory Review of Predictive Models White Paper

Pinnacle Actuarial Resources, Inc. (Pinnacle) is pleased to have the opportunity to provide the following comments in regards to National Association of Insurance Commissioner's (NAIC) Casualty Actuarial and Statistical Task Force (CASTF) second draft of the Regulatory Review of Predictive Models White Paper.

Below are specific comments regarding the second draft of the white paper.

1. Page 3: The first bullet in the description of the GLM states that the target variable follows a probability distribution from the exponential family, but a GLM does not need to be from the exponential family. There are several different distribution family options for a GLM. The phrase "from the exponential family" should be deleted from the bullet.

2. Page 12: "Provide guidance for regulators to determine that individual input characteristics to a model or a sub-model, as well as associated relativities, are not unfairly discriminatory or a "proxy for a protected class."

   The only way this can be done definitively is by either having insurers collect this information or requiring insurance companies to provide detailed data to regulators and letting them add the protected data. Other approaches without having access to actual protected data will have statistical issues.

3. Page 12: "Provide guidance for regulators to identify and minimize unfair discrimination manifested as "disparate impact."

   The phrase "disparate impact" is used often, but is not really defined well anywhere. It will be important for the NAIC to be clear in defining what disparate impact really means. A definition of "disparate impact" should be added to the paper.

4. Page 12: "Provide guidance for regulators that seek a causal or rational explanation why a rating variable is correlated to expected loss or expense, and why that correlation is consistent with the expected direction of the relationship."

   Page 24 (B.3.d): "Obtain a rational explanation for why an increase in each predictor variable should increase or decrease frequency, severity, loss costs, expenses, or any element or characteristic being predicted."

   These items introduce very subjective criteria. A regulator can say that the explanation does not meet the standard or is not reasonable, and does not provide any real recourse for the insurance company. What options will an insurance company have if a regulator thinks a company explanation is not reasonable?

   It can be difficult to prove causation for a correlated variable. It is going to be nearly impossible to define what is acceptable in a clear manner.

5. Page 19 (A.3.c): "Ask for aggregated data (one data set of pre-adjusted/scrubbed data and one data set of post-adjusted/scrubbed data) that allows the regulator to focus on the univariate distributions and compare raw data to adjusted/binned/transformed/etc. data."

   The two data sets being described here are not aggregate data, but more granular data. "Aggregated" should be removed from the statement.

6. Page 34: "In the filed rating plan, be aware of any non-insurance data used as input to the model (customer-provided or other). In order to respond to consumer inquiries, it may be necessary to inquire as to how consumers can verify their data and correct errors."

   What is meant by non-insurance data? Technically, an MVR could be considered as non-insurance data. Does "non-insurance data" mean data external to the insurance company? Non-insurance data should be defined in the paper.

The comments above are the collected comments of the consultants employed or affiliated with Pinnacle. If you have any questions regarding our comments, please contact Laura Maxwell, Pinnacle's Professional Standards Officer, at lmaxwell@pinnacleactuaries.com.

Sincerely,

Laura A. Maxwell, FCAS, MAAA, CSPA
Senior Consulting Actuary

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\Pinnacle Comments

Insurance Services Office, Inc.
545 Washington Boulevard
Jersey City, NJ 07310-1686
www.iso.com

Stephen C. Clarke, CPCU
Vice President
Government Relations
t 201.469.2656
f 201.748.1760
sclarke@iso.com

Kris DeFrain, FCAS, MAAA, CPCU
Director of Research and Actuarial Science
National Assocation of Insurance Commissioners (NAIC) Central Office
1100 Walnut Stree
Suite 1500
Kansas City, MO 64106-2197

re: 6/12/20 Draft White Paper on Best Practices

Dear Ms. DeFrain,

Insurance Services Office, Inc. (ISO) is a countrywide licensed rating/advisory organization serving the property/casualty market. We have extensive experience and expertise in the development of advisory insurance pricing tools including prospective loss costs, rating plans and predictive analytics, including related regulatory issues.

ISO appreciates the opportunity to provide comments on the latest Draft White Paper on Best Practices for Regulatory Review of Predictive Models as published by the CASTF in June 2020. We would like to offer several general comments on the purpose and direction of the best practices document, as well as some specific comments and questions on particular elements within the document.

CASTF has identified 79 information items that should be included in a review of GLM's. The 79 information items appear to go far beyond the aggregation of the current regulatory review practices. We are concerned that the current draft could potentially have the unintended effect of stifling innovation. Given the extensive amount of information being requested some filers may decide that the burden of proof in supporting a GLM is too great and maybe forgo the advantage of a GLM given the cost of complying with the draft best practices.

Here are our detailed comments on the draft.

- On page 3 of the draft, the following sentence refers to the wrong appendix: *"Lastly, provided in this paper are glossary terms (Appendix ~~B~~C) and references."*

- The following statement appears on page 4

  "Though the list of information is long, the insurer should already have internal documentation on the model for more than half of the information listed. The remaining items on the list require either minimal analysis (approximately 25%) or deeper analysis to generate for a regulator (approximately 25%)."

  Can you identity which items the CASTF thinks will require minimal analysis and which will require deeper analysis? This information will be useful to regulators who are concerned with speed to market and want to minimize the additional burden on insurers.

- On page 5 in "Confidentiality" section the following statement is made:

  "State authority, regulations and rules governing confidentiality always apply when a regulator reviews a model used in rating. When NAIC or a third party enters into the review process, the confidential, proprietary, and trade secret protections of the state on behalf of which a review is being performed will continue to apply."

  Can you provide the NAIC legal analysis that concluded that that confidential, proprietary and trade secret protection from the state would apply to NAIC staff that reviews a model?

- B.1.a states "*Identify the type of model underlying the rate filing (e.g. Generalized Linear Model – GLM, decision tree, Bayesian Generalized Linear Model, Gradient-Boosting Machine, neural network, etc.). Understand the model's role in the rating system and provide the reasons why that type of model is an appropriate choice for that role."*

  The information items are intended for GLMs used for personal auto and personal property, so it inconsistent that other model types are mentioned

- B.1.c addresses how validation (hold out) data is used. The GLM paper (Generalized Linear Models for Insurance Rating) that is on the CAS Exam 8 syllabus addresses the use of hold out data. On page 39 it says "Once a final model is chosen, however, we would then go back and rebuild it using *all* of the data, so that the parameter estimates would be at their most credible."

- B.3.b asks for a list of predictor variables considered but not used in the final model and the rationale for their removal. While we appreciate that this is a level 4 item we don't see how the variables not used in a model are relevant to reviewing the filed model. This would be analogous to asking for policy wording considered but not used in a filed policy form.

- C.7.g Obtain a means to calculate the rate charged a consumer.

  While it is feasible for a filer to provide the algorithm with proper trade secret protection, it may not be feasible for a regulator to get all of the input data necessary to produce the model output. Credit and telematics models are an examples of model types where the input data would not be readily available to the regulator,

- On page 43, the following definition of Home Insurance is given "*Home insurance covers damage to the property, contents, and outstanding structures (if applicable), as well as loss of use, liability and medical coverage. The perils covered, the amount of insurance provided, and other policy characteristics are detailed in the policy contract.*[1] "

  The definition should mention that the policy needs to cover a residential dwelling in order for it to be home insurance.

- Appendix B – while we didn't do an exhaustive review of Appendix B on pages 36-40 of the draft we did notice some inconsistencies between Appendix B: Table 1 and Appendix B: Table 2 regarding the mapping of Best Practice Code and Information Element.

  For example, Best Practice Code 2.a in Table 2 references Information Element A.1.a but Best Practice Code 2.a is missing from the mapping in A.1.a in Table 1

A. Selecting Model Input	
A.1. Available Data Sources	
A.1.a	1.b, 1.d, 2.b, 3.a

2. Obtain a clear understanding of the data used to build and validate the model, and thoroughly review all aspects of the model, including assumptions, adjustments, variables, sub-models used as input, and resulting output.		
a. Obtain a clear understanding of how the selected predictive model was built.	2.a	A.1.a, A.2.c, A.2.d, A.2.e, A.2.f, A.3.a, A.3.b, A.4.a, A.4.c, B.1.a, B.1.b, B.1.c, B.1.d, B.1.e, B.1.f, B.1.g, B.1.h, B.1.i, B.1.j, B.2.a, B.2.b, B.2.c, B.2.d, B.2.e, B.2.f, B.3.a, B.3.b, B.3.c, B.3.e, B.4.a, B.4.b, B.4.c, B.4.d, B.4.e, B.4.f, B.4.g, B.4.h, B.4.i, B.4.j, B.5.b, B.5.c, B.6.a, C.1.a, C.4.b, C.4.c, C.5.a

b. Determine that the data used as input to the predictive model is accurate, including a clear understanding how missing values, erroneous values and outliers are handled.	2.b	A.1.a, A.1.b, A.3.a, A.3.b, A.3.c, A.3.d, A.3.e, A.3.f, A.4.a, A.4.b, A.4.c, B.1.h, B.4.d, C.6.a, C.7.h

Another example if inconsistent mappings is Appendix B: Table 1 C.9 and Appendix B: Table 2 4.a, 4.b, 4.c.

C.9. Efficient and Effective Review of a Rate Filing	
C.9.a	4.a
C.9.b	4.a
C.9.c	4.a, 4.b

4. Enable competition and innovation to promote the growth, financial stability, and efficiency of the insurance marketplace.		
a. Enable innovation in the pricing of insurance through acceptance of predictive models, provided they are in compliance with state laws, particularly prohibitions on unfair discrimination.	4.a	C.9.b, C.9.c
b. Protect the confidentiality of filed predictive models and supporting information in accordance with state law.	4.b	C.9.a, C.9.b, C.9.c
c. Review predictive models in a timely manner to enable reasonable speed to market.	4.c	C.9.a, C.9.b, C.9.c

Respectfully Submitted,

Stephen C. Clarke, CPCU

W:\National Meetings\2020\Summer\TF\CasAct\White Paper\July comment letters\ISO comments