Centers for Medicare & Medicaid Services
U.S. Department of Health and Human Services

Center for Consumer Information and Insurance Oversight

HHS Risk Adjustment Data Validation (HHS-RADV) White Paper

December 6, 2019

TABLE OF CONTENTS

GLOSSARY OF TERMS, ACRONYMS, AND DEFINITIONS

Term	Acronym (if applicable)	Definition
Accuracy		The property of being close to a target or true value. It measures how well the sample measurements match the true population value, but does not measure how close sample measurements are to each other (refer to definition of 'precision' below).
Bootstrap Resampling Technique		A method of testing precision, wherein one large sample is taken from the parent distribution and then multiple samples with replacement from that initial sample are drawn. It is used to determine standard errors and confidence intervals when the underlying distribution is unknown, when sample sizes may be too small, and/or when no formula may exist for a complex calculation.
Center for Consumer Information and Insurance Oversight	CCIIO	The component within CMS charged with helping implement many reforms of the Patient Protection and Affordable Care Act (PPACA). CCIIO oversees the implementation of the provisions related to private health insurance. In particular, CCIIO works with states on the operation of the Health Insurance Exchanges. CCIIO works closely with state regulators, consumers, and other stakeholders to ensure the PPACA best serves the American people.
Centers for Medicare & Medicaid Services	CMS	A federal agency within the United States Department of Health and Human Services that administers the Medicare program and works in partnership with state governments to administer Medicaid, the State Children's Health Insurance Program, and the PPACA private market reforms.
Confidence Level		In terms of HHS-RADV error estimation, the theoretical probability that an issuer whose population-level failure rate for an HCC group is very similar to the national mean will *not* be found to be an outlier, given that all statistical assumptions about the underlying distribution are upheld.
Default Data Validation Charge	DDVC	Charge imposed under 45 CFR § 153.630(b)(10) if an issuer fails to engage an initial validation auditor or submit initial validation audit (IVA) results. The DDVC is calculated similarly to the Risk Adjustment Default Charge (RADC), but is an independently calculated and assessed penalty.
Demographic and Enrollment	D&E	Describes an enrollee's demographics and enrollment status.
External Data Gathering Environment	EDGE	Issuer-distributed data collection services (also known as an EDGE server) that issuers in states where HHS operates a risk adjustment program are required to establish and to compile enrollment, pharmaceutical claims and medical claims information on enrollees in risk adjustment covered plans from issuers' proprietary systems. An EDGE server runs HHS-developed software designed to verify submitted data, execute risk adjustment processes, and generate summary reports for submission to HHS.

Term	Acronym (if applicable)	Definition
Error Rate		The rate at which an outlier issuer's risk score is adjusted based on HHS-RADV results. If an issuer is identified as a group failure rate outlier in one or more hierarchical condition categories (HCC) groups, its individual enrollee risk scores are adjusted based on the differences between the issuer's group failure rate and the national mean HCC group failure rate in every HCC group in which the issuer is identified as an outlier. The issuer's error rate equals $1 - \frac{stratum\ weighted\ sum\ of\ adjusted\ enrollee\ risk\ scores}{sum\ of\ original\ EDGE\ enrollee\ risk\ scores}$
Failure Rate		The rate at which the frequency of HCCs identified through the IVA or second validation audit (SVA) differ from the frequency of the HCCs identified on EDGE. Failure rate equals $1 - \frac{HCC\ frequency\ in\ IVA\ or\ SVA}{HCC\ frequency\ on\ EDGE}$ During HHS-RADV error estimation, the HCC failure rate is calculated for each HCC to determine the low, medium and high HCC groups that determine the national mean and confidence intervals for each HCC group. Then, each issuer's failure rate is calculated for each HCC group to determine whether the issuer is a group failure rate outlier, which would lead to a non-zero error rate and an adjustment to the issuer's enrollee risk scores.
Failure Rate Z Score		An issuer's HCC group failure rate compared to the weighted mean of that HCC grouping measured in weighted standard deviations from the mean. It is a measure of how many standard deviations below or above the failure rate mean an issuer's failure rate is within an HCC group.
Finite Population Correction	FPC	A formula that assists in determining a modified sample size for issuers with fewer than 4,000 enrollees. It is used to define both the standard error of the mean and the standard error of the proportion.
Group Adjustment Factor		The adjustment factor calculated for each HCC group, assigned when the issuer's HCC group failure rate is outside of the upper or lower boundary of an HCC group in HHS-RADV. This factor is weighted and used to compute the issuer's error rate and the enrollee-level adjusted risk score(s).
Health and Human Services	HHS	The federal government department whose mission is to enhance and protect the health and well-being of all Americans. HHS fulfills its mission by providing for effective health and human services and fostering advances in medicine, public health, and social services.
Health and Human Services' Risk Adjustment Data Validation	HHS-RADV	The data validation process that is part of the HHS-operated risk adjustment program (HHS-RA) under section 1343 of PPACA. The process involves validating a statistically valid sample of enrollment and health status data submitted by issuers of risk adjustment covered plans to their respective EDGE servers.
Health Insurance Oversight System	HIOS	A system created to facilitate several types of data collections from the Departments of Insurance for states/territories as well as health insurance issuers that sell health insurance coverage. HIOS collects insurance company and product information, such as the issuer names, addresses, contact information, and product level data. Each issuer entity is assigned a unique HIOS ID by state.

Term	Acronym (if applicable)	Definition
Hierarchical Condition Categories	HCC	A payment model that uses coding to identify health conditions documented by health professionals and assigns a risk score factor. HHS-operated risk adjustment uses HCCs to estimate a risk score for each enrollee in an issuer's risk adjustment population and uses those risk scores to calculate the issuer's plan liability risk score (PLRS). The PLRS is used in the HHS risk adjustment state payment transfer formula. Similar HCCs are placed in a hierarchy and are grouped together in the HHS-operated risk adjustment model (See Appendix D for more information).
HCC Groupings or HCC Groups		In HHS-RADV, all HHS-RA HCCs are grouped into high, medium, and low groups based on individual HCC failure rates across all issuers. A confidence interval is calculated for each HCC grouping at the national level. If an individual issuer's failure rate for at least one HCC grouping is outside the confidence interval for that HCC grouping, the issuer is determined to be an outlier in HHS-RADV.
Initial Validation Audit	IVA	The initial validation audit of enrollment data, claims data and health status data submitted by the issuer to HHS for risk adjustment covered plans. This audit is conducted by an independent audit entity (IVA Entity) hired by the issuer. Findings from the IVA must be submitted to HHS for review during the SVA.
Initial Validation Audit Entity	IVA Entity	The independent audit entity contracted by the issuer to conduct the IVA.
Issuer		A licensed entity offering health insurance coverage within a state. Each issuer entity is assigned a unique HIOS ID by state.
Lower Bound		The lower limit of a confidence interval, used in HHS-RADV outlier identification and in measuring precision.
Medicare Advantage Risk Adjustment Data Validation	MA-RADV	The risk adjustment data validation program that applies to Medicare Advantage plans participating in Medicare Part C under the Social Security Act.
Neyman Allocation		The statistical method that calculates the optimal number to be sampled from each stratum, proportional to each stratum's contribution to the total standard deviation of the population (i.e., more variable strata should be sampled more intensely).
Outlier		A value that falls outside of an established threshold. In HHS-RADV, a HIOS ID with a failure rate that falls outside of the HCC Group upper or lower boundary is an outlier. A HIOS ID may be identified as an outlier in one, two, or all three HCC Groups.
Pairwise Means Test		A hypothesis-testing procedure to determine if two population means are different when there is a one-to-one correspondence between the values in the two samples.
Patient Protection and Affordable Care Act	PPACA	Reforms certain aspects of the private health insurance industry and public health insurance programs, including increasing insurance coverage of pre-existing conditions and expanding access to insurance to Americans. The PPACA (Pub. L. 111–148) was enacted on March 23, 2010. The Health Care and Education Reconciliation Act of 2010 (Pub. L. 111–152), which amended and revised several provisions of the PPACA, was enacted on March 30, 2010. In this white paper, we refer to the two statutes collectively as the "Patient Protection and Affordable Care Act" or "PPACA".
Payment Notice		HHS's annual rulemaking that establishes the parameters and policies governing health insurance coverage for the upcoming benefit year, formally called the HHS Notice of Benefit and Payment Parameters.

Term	Acronym (if applicable)	Definition
Payment Error Rate Measurement	PERM	The Payment Error Rate Measurement program measures and reports a national improper payment rate for Medicaid and the Children's Health Insurance Program.
Plan Liability Risk Score	PLRS	The enrollment-weighted average risk score of all enrollees in a particular risk adjustment-covered plan for an issuer.
Practical Confidence Level		The simulated, empirical probability that an issuer whose population-level failure rate for an HCC group is very similar to the national mean will *not* be found to be an outlier given possible violations to statistical assumptions about the underlying distribution that may be present in actual HHS-RADV data (refer to the definition of "Confidence Level" above).
Precision		A measurement of how close in value sampled observations are likely to be to one another. It refers to the dispersion of a set of observations, and does not measure how closely sample observations match the true population (refer to the definition of "accuracy" above).
Risk Adjustment	RA or HHS-RA	A premium stabilization program established by the PPACA. The overall goal of risk adjustment is to eliminate premium differences among plans based solely on favorable or unfavorable risk selection in the individual and small group markets both inside and outside of the Exchanges. Risk adjustment accomplishes this by transferring funds from issuers with lower risk enrollees to issuers with higher risk enrollees.
Risk Adjustment Default Charge	RADC	Charge imposed under 45 CFR § 153.740(b) if an issuer of a risk adjustment covered plan fails to establish an EDGE server or fails to provide HHS with access to the required data on the EDGE server, such that HHS cannot apply the federally certified risk adjustment methodology.
Risk Adjustment Prescription Drug Category	RXC	The use of a drug to impute a diagnosis (or indicate the severity of the diagnosis) otherwise indicated through medical coding in a hybrid diagnoses-and-drugs risk adjustment model. Beginning with the 2018 benefit year, RXCs are utilized in the HHS-RA program to calculate an adult enrollee's risk score. As a result, IVA Entities are required to validate the RXCs of sampled enrollees beginning with the 2018 benefit year of HHS-RADV.
Second Validation Audit	SVA	The independent, third-party audit of the IVA Entity's Audit results performed by the SVA Entity.
Second Validation Audit Entity	SVA Entity	The entity retained by HHS to validate the IVA findings.
Standard Deviation	SD	The measurement of the amount of variability, or dispersion, for a set of selected data values. The standard deviation is equal to the square root of the variance.
Standard Error	SE	An estimate of the standard deviation of a sampling distribution. A measure of the variability of a statistic.
Strata		The subsets of a population being sampled. In HHS-RA, these are mutually exclusive groups within the population and are constructed based on recorded risk score, age, and presence of HCCs (or RXCs).
Upper Bound		The upper limit of a confidence interval, used in HHS-RADV outlier identification and in measuring precision.

EXECUTIVE SUMMARY

Section 1343 of the Patient Protection and Affordable Care Act (PPACA) established a risk adjustment program to provide payments to health insurance issuers that attract high-risk enrollees, such as those with chronic conditions, to reduce the incentives for issuers to avoid those enrollees, and to lessen the potential influence of risk selection on the premiums that issuers charge.[1] Risk adjustment (RA) is an essential component for markets that require guaranteed issue and community rating to protect issuers from adverse risk selection and create incentives for issuers to offer a wide range of plan designs that are particularly valuable to sicker individuals. The risk adjustment program authorized under Section 1343 of PPACA is the only permanent premium stabilization program under PPACA and it applies to non-grandfathered plans in the individual and small group (or merged) markets both inside and outside of the Exchanges. Consistent with section 1321(c)(1) of PPACA[2], the Department of Health and Human Services (HHS) is responsible for operating the program on behalf of any states that do not elect to do so. Prior to the 2017 benefit year, all states and the District of Columbia, except Massachusetts, participated in the HHS-operated risk adjustment program (HHS-RA) and since the 2017 benefit year, all states and the District of Columbia have participated in the HHS-operated risk adjustment program. The HHS-operated risk adjustment program results in the transfer of billions of dollars among health insurance issuers in individual, small group, catastrophic, and merged market risk pools annually.

To ensure the integrity of the risk adjustment program, the Centers for Medicare & Medicaid Services (CMS), on behalf of HHS, performs risk adjustment data validation, also known as HHS risk adjustment data validation (HHS-RADV). One of the primary purposes of HHS-RADV is to validate the accuracy of data submitted by issuers for the purposes of risk adjustment transfer calculations. HHS-RADV serves as an audit of the information used in establishing an enrollee's risk score for purposes of calculating the issuer's plan liability risk score (PLRS) under the risk adjustment program. The findings from HHS-RADV are used to adjust issuers' enrollee risk scores and risk adjustment transfers. Error estimation is the multi-step process of using the HHS-RADV findings to calculate the adjustment to issuers' risk scores and risk adjustment transfers. Due to the budget-neutral nature of the HHS-operated risk adjustment program, adjustments to one issuer's enrollee risk scores and risk adjustment transfers based on HHS-RADV findings will affect all other issuers in the state market risk pool.

The purpose of this white paper is to outline and seek feedback on certain HHS-RADV issues that we may use to inform future HHS-RADV policy. Since we began developing HHS-RADV in 2013, we have sought feedback from stakeholders in its design and operation. We conducted two pilot years of HHS-RADV for the 2015 and 2016 benefit years[3] before applying

---

[1] 42 USC 18063.
[2] 42 USC 18041(c)(1).
[3] HHS-RADV was not conducted on 2014 benefit year data. See FAQ ID 11290a (March 7, 2016) available at: https://www.regtap.info/faq_viewu.php?id=11290.

2017 benefit year HHS-RADV findings to adjust risk scores used in risk adjustment transfers for the 2018 benefit year.[4,5] Based on our experience from these initial years of conducting HHS-RADV and analysis of currently available information, HHS is considering potential modifications to certain aspects of the HHS-RADV program for future benefit years.

Specifically, HHS is considering potential modifications to four specific aspects of the HHS-RADV program: 1) enrollee sampling; 2) outlier detection; 3) the error rate calculation and 4) the application of HHS-RADV results, as defined below.

- Enrollee sampling: is the method by which a statistically valid sample of enrollees for each issuer is selected for validation of their risk scores in HHS-RADV.[6] This white paper considers whether the current enrollee sampling methodology, which is based on Medicare Advantage Risk Adjustment Data Validation (MA-RADV) error rates and results in a sample size of 200 enrollees for most issuers, should be adjusted.
- Outlier detection: is the process by which HHS uses all issuers' HHS-RADV results to establish national metrics (e.g. means and confidence intervals) to determine whether an issuer's rate of failure to validate its enrollees' risk scores at the hierarchical condition category (HCC) level is outside of an acceptable range of variation (an outlier). Based on our experiences in the initial years of HHS-RADV, this white paper assesses the sensitivity of the current outlier detection methodology and considers options to modify the outlier detection process to more precisely identify true outliers. This white paper also discusses the influence of HCC hierarchies in outlier detection.
- Error rate calculation: is the calculation of the percentage by which an outlier issuer's risk score is adjusted based on the issuer's failure to validate the HCCs associated with enrollees selected for audit. This white paper examines alternatives to the current methodology that determines an outlier issuer's risk score adjustment by calculating the difference between the issuer's HCC group failure rate and the weighted mean group failure rates from the national metrics. Specifically, this white paper focuses on alternative options to address cases where the outlier issuer may have a failure rate that is only slightly outside of the acceptable range of variation, as well as cases where an outlier issuer has a negative failure rate.
- Application of HHS-RADV results: are done using a prospective approach. Currently, HHS uses an issuer's HHS-RADV error rate to adjust the issuer's average risk score and risk adjustment transfer amount in the transfer year following the HHS-RADV

[4] The one exception is for issuers who exited all markets in the state for the 2018 benefit year. For these issuers, their 2017 benefit year HHS-RADV results applied to their respective 2017 benefit year PLRS and were used to adjust 2017 benefit year risk adjustment transfer amounts in the applicable state market risk pool.

[5] HHS does not calculate risk adjustment transfers for state market risk pools in which there is only one issuer (sole market risk pool issuers) and those issuers are not required to conduct HHS-RADV for that state market risk pool for the applicable benefit year. See the 2020 Payment Notice, 84 FR at 17504. Also see the 2019 Payment Notice; Final Rule; 83 FR 16930 at 16967.

[6] The validation rate of these enrollees' risk scores is also used in error estimation to calculate an outlier issuer's error rate. This error rate is applied to adjust its risk scores and the issuers' risk adjustment transfers in the applicable state market risk pool.

result (e.g., 2018 benefit year HHS-RADV error rates are generally applied to 2019 risk scores and risk adjustment transfers).[7] This white paper considers a change to the application of HHS-RADV results to better reflect actuarial risk by applying HHS-RADV results to the benefit year being audited (e.g., 2021 benefit year HHS-RADV error rates could be applied to 2021 benefit year risk scores and risk adjustment transfers).

The options in this white paper were developed based on HHS's ongoing internal analysis of potential refinements to the HHS-RADV program for future benefit years, as well as comments received on HHS-RADV through notice-and-comment rulemaking and through listening sessions with stakeholders. We are seeking comments on the options outlined in this white paper to help inform potential future rulemaking in these areas. Commenters should submit comments by Monday, January 6, 2020 to CCIIOACARADataValidation@cms.hhs.gov with the subject line of "December 2019 HHS-RADV White Paper."

[7] The exception to this general rule is for exiting issuers. See supra note 4.

1. HHS Risk Adjustment Data Validation (HHS-RADV) Overview

1.1 Purpose and Structure of this White Paper

Based on our experience from the two pilot years and the first payment year of HHS-RADV, HHS is examining modifications to certain aspects of the HHS-RADV program. In particular, the purpose of this white paper is to outline potential options to modify the methodology for enrollee sampling, amend the current process that determines whether an issuer is an outlier, alter the error rate calculation that determines outlier issuers' risk score adjustments, and change the benefit year application of HHS-RADV results.

Chapter 1: The first chapter of this white paper provides an overview of HHS-RADV.

Chapter 2: The second chapter of this white paper discusses potential options to modify the current enrollee sampling methodology.

Chapter 3: The third chapter of this white paper outlines potential modifications to the current process for determining whether an issuer is an outlier.

Chapter 4: The fourth chapter of this white paper discusses potential options to revise the current calculation of an outlier issuer's error rate.

Chapter 5: The fifth chapter of this white paper considers changing the application of HHS-RADV results from a prospective approach to align with the benefit year being audited.

We developed this white paper for comment based on our internal analysis of HHS-RADV results and comments received regarding stakeholders' experiences with the initial years of HHS-RADV. Over the course of July and August of 2019, CMS also conducted a series of stakeholder engagement sessions about the initial years of HHS-RADV to hear what modifications may be needed for future benefit years. Those stakeholder discussions helped inform the policy issues considered in this white paper.

This white paper does not address operational issues that may occur during HHS-RADV, such as medical record retrieval issues or national provider coding standards. Those issues are addressed annually at the "Lessons-Learned Meeting" with Initial Validation Audit Entities (IVA Entities) that is hosted by CMS and through operational user group calls. Guidance on these operational issues is largely provided through the HHS-RADV Protocols that are published annually.

1.2 Statutory and Regulatory Background of HHS-RADV

Section 1343 of the PPACA provides for a permanent risk adjustment program for non-grandfathered plans in the individual and small group markets, both inside and outside of the Exchanges. The PPACA directs the Secretary, in consultation with the states, to establish criteria and methods to be used in carrying out risk adjustment activities, such as determining the

actuarial risk of plans within a state market risk pool.[8] The statute also provides that the Secretary may utilize criteria and methods similar to the ones utilized under Medicare Parts C or D.[9] States electing to operate a risk adjustment program, or HHS on behalf of states not electing to operate a risk adjustment program, assess charges to issuers with plans that experience lower than average actuarial risk and use the collected charges to pay issuers with plans that have higher-than-average actuarial risk. For the 2014-2016 benefit years, all states and the District of Columbia, except Massachusetts, participated in the HHS-operated risk adjustment program. Since the 2017 benefit year, all states and the District of Columbia have participated in the HHS-operated risk adjustment program.

The risk adjustment program is designed to facilitate a plan enrolling a higher proportion of high-risk enrollees charging the same average premium (other factors being equal) as a plan enrolling a higher proportion of low-risk enrollees, shifting the focus of plan competition to quality, efficiency, and value. Risk adjustment accomplishes this goal by transferring funds from issuers with lower risk enrollees to issuers with higher risk enrollees. The HHS-operated program calculates a plan average risk score for each covered plan based upon the relative risk of the plan's enrollees, and applies a state payment transfer formula in order to determine risk adjustment payments and charges between plans within a state market risk pool. Beginning with the 2018 benefit year, the program includes a high-cost risk pool, which helps ensure that risk adjustment transfers better reflect the average actuarial risk in a state market risk pool.[10] The HHS-operated risk adjustment program results in billions of dollars being transferred among health insurance issuers in individual, small group, catastrophic, and merged market risk pools annually. To ensure these funds are transferred appropriately, program integrity is an integral part of the risk adjustment program.[11]

To ensure the integrity of the risk adjustment program, CMS, on behalf of HHS, performs risk adjustment data validation, also known as HHS-RADV, to validate the accuracy of data submitted by issuers for the purposes of risk adjustment transfer calculations. HHS-RADV ensures that transfers reflect issuers' actual actuarial risk and that risk adjustment assesses charges to issuers with plans with lower-than-average actuarial risk while making payments to issuers with plans with higher-than-average actuarial risk. Thus, the purpose of HHS-RADV is to promote confidence and stability in the budget-neutral transfer methodology used by the HHS-operated risk adjustment program by ensuring the integrity and quality of data provided from issuers. The priorities in implementing HHS-RADV are to promote consistency and a level playing field by establishing uniform audit requirements, and to protect private information by limiting data transfers during the data validation process. HHS believes that a robust HHS-

[8] 42 USC 18063(b).
[9] Ibid.
[10] High-cost risk pool transfers are not subject to HHS-RADV.
[11] HHS also has general audit authority over issuers of risk adjustment covered plans pursuant to 45 C.F.R. § 153.620(c).

RADV process is critical to ensuring issuer confidence and to meeting the goals of the risk adjustment program.

To initially develop the HHS-RADV process, we sought the input of issuers, consumer advocates, providers, and other stakeholders. We issued the "Affordable Care Act HHS-Operated Risk Adjustment Data Validation Process White Paper" on June 22, 2013 (the 2013 white paper).[12] The 2013 white paper discussed and sought comments on a number of potential considerations for the development and operation of the HHS-RADV program. Based on feedback that we received on the 2013 white paper, we promulgated regulations to implement HHS-RADV that we have modified over the years.[13]

45 C.F.R § 153.350(a) requires the state, or HHS on behalf of the State, to ensure proper validation of a statistically valid sample of risk adjustment data from each issuer that offers at least one risk adjustment covered plan in that State. Specifically, for the HHS-operated risk adjustment program, 45 C.F.R. § 153.630 requires an issuer of a risk adjustment covered plan[14] in a state where HHS is operating risk adjustment to have an initial and second validation audit performed on its risk adjustment data for the applicable benefit year. Each issuer must engage an independent validation auditor to perform the initial validation audit (IVA) of a sample of risk adjustment data selected by HHS. After the IVA Entity has validated the HHS-selected sample, a subsample of that sample is also validated in a second validation audit (SVA). The SVA is conducted by an entity HHS retains to verify the accuracy of the findings of the IVAs. 45 C.F.R. § 153.350 also allows the state, or HHS on behalf of the State, to adjust the plan average actuarial risk for a risk adjustment covered plan based on errors discovered as a result of data validation and to use those errors discovered in data validation to adjust charges and payments to all risk adjustment covered plans based on the adjustment to the plan average actuarial risk from errors. Lastly, 45 C.F.R. § 153.350(d) requires the State, or HHS on behalf of the State, to establish the processes for an issuer to dispute and appeal its HHS-RADV results.[15]

To operationalize HHS-RA, each issuer is required to have an External Data Gathering Environment (EDGE) server on which the issuer must submit masked enrollee demographics, claims, and encounter diagnosis-level data in a format specified by HHS. HHS queries these EDGE servers, directing issuers to execute software on their respective EDGE servers to generate summary reports that HHS uses to calculate the enrollee-level risk score for the purpose of determining the average PLRS for each state market risk pool, as well as individual issuers' PLRSs. The difference between issuers' PLRSs and the average PLRS is used to calculate the transfers for issuers of risk adjustment covered plans within a state market risk pool. To ensure the integrity of this process, HHS-RADV serves as an audit of the information derived from the

---

[12] A copy of the Affordable Care Act HHS-Operated Risk Adjustment Data Validation Process White Paper (June 22, 2013) is available at: https://www.regtap.info/uploads/library/ACA_HHS_OperatedRADVWhitePaper_062213_5CR_050718.pdf.

[13] See, e.g., 45 C.F.R. §§ 153.350 and 153.630. An overview of the specific modifications made to the HHS-RADV regulations over the years appears in Appendix A.

[14] See 45 C.F.R. § 153.20 for a definition of "risk adjustment covered plan."

[15] See 45 C.F.R. § 153.630(d).

demographic, claims and diagnosis data submitted to the issuers' EDGE servers for use in establishing an enrollee's risk score for purposes of calculating the issuer's PLRS. Therefore, the statistically valid enrollee sample is derived from the information on the issuer's EDGE server.

To operationalize HHS-RADV, HHS modeled many aspects of HHS-RADV processes after the Medicare Advantage risk adjustment data validation (MA-RADV) program. For example, HHS elected to adopt medical records as the authoritative source to verify diagnoses from the EDGE server, and requires that certified medical coders perform medical record reviews. Because HHS's risk adjustment methodology uses a more comprehensive set of data elements than Medicare Advantage, the HHS-RADV data collection approach is more robust, and HHS's data validation approach is broader for HHS-RADV.

HHS conducted two pilot years of HHS-RADV for the 2015 and 2016 benefit years to give HHS and issuers experience with how the audits would be conducted prior to applying HHS-RADV results to adjust issuers' risk scores and risk adjustment transfers in the applicable state market risk pool. The 2017 benefit year HHS-RADV was the first non-pilot year, and resulted in adjustments to issuers' risk scores and risk adjustment transfers as a result of HHS-RADV findings.[16]

1.2.1 HHS-RADV Process Overview

HHS-RADV is a six step process. The first step in the HHS-RADV process is the selection of a sample of an issuer's enrollees whose risk adjustment data from the EDGE server will be validated by the IVA Entity. For the audit, HHS applies a sampling methodology to choose a statistically valid sample of enrollees based on the enrollee-level risk score distributions for each issuer. HHS designed the sampling methodology to ensure that the sample covers critical subpopulations of enrollees for each issuer by dividing each issuer's population into 10 strata, representing different age and risk score bands, and sampling from each stratum. Based on sample size precision analyses and calculations using proxy error rate data from the MA-RADV program, issuers of sufficient size currently have a sample size of 200 enrollees across all state market risk pools and risk adjustment covered plans. The second step of the HHS-RADV process is the IVA. The issuer must ensure that its selected IVA Entity is reasonably capable of performing this task, and is reasonably free of conflicts of interest, and is therefore able to conduct the IVA in an impartial manner. HHS expects issuers to ensure that the IVA is conducted in the following manner:

- The issuer provides the IVA Entity with enrollment, claims, and medical record documentation to validate issuer-submitted risk adjustment data for each enrollee in the sample;
- The issuer and IVA Entity determine a timeline and information transfer methodology that satisfies data security and privacy requirements and enables the IVA Entity to meet HHS-established timelines; and

[16] The Summary Report of 2017 Benefit Year HHS-RADV Adjustments to Risk Adjustment Transfers released on August 1, 2019 is available at: https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/BY2017-HHSRADV-Adjustments-to-RA-Transfers-Summary-Report.pdf.

- The IVA Entity validates the data of each enrollee in the sample in accordance with the standards established by HHS.

Once these steps are completed, the IVA Entity provides HHS with the final results from the IVA and all requested information for HHS to complete the SVA.

Under the third step of HHS-RADV, HHS retains an SVA Entity to conduct the SVA to verify the accuracy of the findings of the IVA. HHS selects a subsample of the IVA sample of enrollees for review. When the SVA Entity performs the data validation audit of the enrollee subsample, the SVA Entity adheres to the same audit standards applicable to the IVA, but only reviews enrollee information that was submitted to HHS at the conclusion of the IVA by the issuer and the IVA Entity.

HHS selects a small subsample of enrollees for the SVA Entity review using a sampling methodology that allows for pairwise means testing to detect any statistical difference between the initial and second validation audit results. If the pairwise means test results suggest that the difference in enrollee results between the IVA and SVA is not statistically significant, HHS uses the IVA results for error estimation and calculation of adjustments to the issuer's PLRS, if the issuer is determined to be an outlier. If the pairwise means test results suggest a statistical difference based on the initial SVA sample, the SVA Entity would perform another validation audit on a larger subsample of the enrollees previously subject to the IVA. HHS then repeats pairwise means testing. If a statistical difference is still found between the IVA and the SVA of the larger subsample, the SVA sample is expanded to larger subsample sizes with pairwise means testing repeated for each such expansion until the full SVA subsample of 100 enrollees is reviewed. If a statistical difference is still found between the IVA and the expanded SVA sample(s) (up to 100 enrollees), HHS will use the SVA results for error estimation and calculation of adjustments to the issuer's PLRS, rather than the IVA results, if the issuer is determined to be an outlier.

The fourth step in the HHS-RADV process is error estimation. Using the relevant validation audit data determined in the prior step (i.e., either the IVA or SVA findings, as applicable), HHS derives issuer-level failure rates for each HCC. The HCC failure rate represents the rate at which the EDGE HCC cannot be validated during the HHS-RADV process. Then, HHS aggregates all issuers' failure rates and creates HCC groupings, national means, and 95 percent confidence intervals at 1.96 standard deviations for that benefit year of HHS-RADV. Under this process, each HCC used in the risk adjustment program for that benefit year is organized into three HCC groupings, high, medium and low, based on the individual HCC's failure rate across all participating issuers' HHS-RADV results. The aggregated failure rates from each of these HCC groupings are then used to create the national means and confidence intervals for each HCC group. These national means and confidence intervals determine whether an issuer is an outlier for an HCC grouping. If an issuer's HCC group failure rate is outside the national confidence intervals in any of three HCC groupings (low, medium, or high groupings), the issuer is determined to be an outlier. If the issuer is an outlier, its failure rates are used to calculate its error rate, which adjusts its risk score to reflect the inaccuracy of the risk score calculated during risk adjustment. If the issuer is not an outlier, the issuer is determined to have a zero error rate

result for HHS-RADV for that benefit year and its risk score remains unadjusted. Later in this paper, we describe the error estimation process in more detail.

The fifth step of HHS-RADV is the annual discrepancy and administrative appeals processes. An issuer is required to confirm the information provided by HHS or file a discrepancy within a certain number of days of notification by HHS of certain HHS-RADV results. The discrepancy and appeal processes apply to the IVA sample, the findings of the SVA, and the calculation of the risk score error rate.[17]

The sixth and final step of HHS-RADV is adjusting risk adjustment transfer amounts to reflect the level of inaccuracy determined in the fourth step of error estimation. Except for exiting issuers,[18] when an issuer is identified as an outlier, the issuer's error rate is used to amend the issuer's subsequent benefit year risk score that is used to calculate their risk adjustment transfer on a prospective basis, e.g. 2017 benefit year HHS-RADV results amended 2018 benefit year risk adjustment risk scores and 2018 benefit year risk adjustment transfers. Because risk adjustment is budget neutral, a change in risk score for one issuer in a state market risk pool affects the statewide average risk score for that state market risk pool, impacting all other issuers in the state market risk pool. These changes in risk scores are then applied to adjust risk adjustment transfers for the applicable state market risk pool, and the adjustments are collected and distributed two years later.

1.2.2 Original HHS-RADV Error Estimation Methodology

The original HHS-RADV error estimation methodology was finalized in the 2015 Payment Notice.[19] Under the original methodology, HHS would use the results of the IVA or SVA, as applicable, as the basis for calculating a corrected risk score for each risk score for each enrollee in the issuer's sample population. Under this methodology, the majority of issuers would have a HHS-RADV adjustment since any failure to validate an HCC has the potential to result in an adjustment (see Appendix B for more detail). As a result, almost all issuers for a given benefit year would have seen a change in risk adjustment transfers due to HHS-RADV findings.

1.2.3 Current HHS-RADV Error Estimation Methodology

In the 2019 Payment Notice, HHS explained that we believe that some variation and error should be expected in the compilation of data for risk scores because providers' documentation of enrollee health status varies across provider types and groups.[20] Our experiences with the MA-RADV program and the HHS-RADV pilot for the 2015 benefit year reinforced this belief. Thus, to avoid adjusting all issuers' risk adjustment transfers for expected variation and error in EDGE

---

[17] Issuers cannot appeal the results of the IVA as the IVA Entity is under contract with the issuer and HHS does not produce the IVA results. See, e.g., the HHS Notice of Benefit and Payment Parameters for 2018; Final Rule; 81 FR 94056 at 94106 (December 22, 2016).

[18] For exiting issuers, their amended risk scores are applied to the prior year's risk adjustment transfer amounts for the applicable state market risk pool, e.g., exiting issuer 2017 benefit year HHS-RADV results amended 2017 benefit year risk scores, which were applied to 2017 benefit year risk adjustment transfers for the applicable state market risk pools.

[19] 79 FR 13743 at 13755-13770.

[20] See 83 FR 16930 at 16961-16965.

HCCs, we adopted the current methodology that evaluates material statistical deviation in failure rates.

Under the current methodology, HHS amends an issuer's risk score only when the issuer's failure rate materially deviates from a statistically meaningful national value. HHS determines the national statistically meaningful value as the weighted mean failure rate calculated based on all issuers' HHS-RADV results. As previously described, to apply this methodology, HHS uses the failure rates for each HCC to group each HCC into three HCC groupings. These HCC groupings are determined by first ranking all HCC failure rates and then dividing the rankings into three groupings weighted by total observations of that HCC across all issuers' IVA samples, assigning each HCC into a high, medium, or low HCC grouping. We calculate an issuer's HCC group failure rate as:

$$GFR_i^G = 1 - \frac{Freq\_IVA_i^G}{Freq\_EDGE_i^G}$$

Where:

$Freq\_EDGE_i^G$ is the number of HCCs in group *G* in the EDGE sample of issuer *i*.

$Freq\_IVA_i^G$ is the number of HCCs in group *G* in the IVA sample of issuer *i*.

GFR_i^G is *i*'s group failure rate for the HCC group *G*.

We will also calculate the weighted mean failure rate and the standard deviation of each HCC group as:

$$\mu^*(GFR^G) = 1 - \frac{\sum Freq\_IVA_i^G}{\sum Freq\_EDGE_i^G}$$

$$Sd(GFR^G) = \sqrt{\frac{\sum_i Freq\_EDGE_i^G * \left(GFR_i^G - \mu(GFR^G)\right)^2}{\sum_i Freq\_EDGE_i^G}}$$

Where:

$\mu(GFR^G)$ is the weighted mean of GFR_i^G of all issuers for the HCC group *G* weighted by all issuers' sample observations in each group.

$Sd(GFR^G)$ is the weighted standard deviation of GFR_i^G of all issuers for the HCC group *G*.

The issuer's HCC group failure rates are than compared against the national metrics for each HCC grouping. If an issuer's failure rate for an HCC group falls outside of the 95 percent confidence interval with a 1.96 standard deviation cutoff calculated based on the weighted mean

failure rate for the HCC group, the failure rate for the issuer's HCCs in that group is considered an outlier. If all issuers' HCC group failure rates in a state market risk pool do not materially deviate from the national mean of failure rates (that is, no issuers in a state market risk pool are outliers), we do not apply any adjustments to issuers' risk scores for that benefit year in the respective state market risk pool.

Under the current methodology, when an issuer is determined to be an outlier, the adjustment to an enrollee's total risk score is calculated as the ratio of the total amended risk score for individual HCCs to the total risk score components for individual HCCs submitted to the EDGE server for the enrollee. For example, if an issuer has one enrollee with the HIV/AIDS HCC and the issuer's HCC group adjustment rate is 10 percent (the difference between the issuer's group failure rate and the weighted mean group failure rate) for the HCC group that contains the HIV/AIDS HCC, the enrollee's HIV/AIDS HCC risk score coefficient would be reduced by 10 percent. For each enrollee, we calculate the total amended risk score across all outlier HCCs as:

$$AdjRS_{i,e} = EdgeRS_{i,e} * (1 - Adjustment_{i,e})$$

Where:

$EdgeRS_{i,e}$ is the risk score for EDGE HCCs of enrollee *e* of issuer *i*.

$AdjRS_{i,e}$ is the amended risk score for sampled enrollee *e* of issuer *i*.

$Adjustment_{i,e}$ is the adjustment factor by which we estimate the EDGE risk score exceeds or falls short of the initial or second validation audit projected risk score across all HCCs and HCC groups for sampled enrollee *e* of issuer *i.*

We then calculate an issuer's risk score error rate using the EDGE risk score and amended risk score for all enrollees in the sample. The current methodology for extrapolating amended risk scores from the sample to the population and determining the issuer's risk score error rate is consistent with the approach under the original methodology. CMS obtains the weight in the error rate calculation formula by multiplying the ratio of an enrollee's stratum size and the issuer's population size to the total number of sample enrollees that are in the same stratum as the enrollee. The formula to compute the risk score error rate using the stratum-weighted risk score for issuer *i* before and after the adjustment is shown as:

$$ErrorRate_i = 1 - \frac{\sum_e (w_e * AdjRS_{i,e})}{\sum_e (w_e * EdgeRS_{i,e})}$$

Where:

$$w_e = \frac{stratum\ size\ in\ population}{number\ of sample\ enrollees\ of\ the\ stratum}$$

We then apply the risk score error rate to the prospective benefit year's calculated PLRS and risk adjustment transfers.[21] The current methodology results in fewer state market risk pools and issuers receiving amendments to their risk scores and risk adjustment transfers as a result of HHS-RADV findings than the original methodology.[22] The current methodology applied beginning with the 2017 benefit year HHS-RADV.[23]

1.3 HHS-RADV EXPERIENCE

As previously mentioned, the 2015 and 2016 benefit years were pilot years for HHS-RADV. The 2017 benefit year was the first year in which risk adjustment transfers were adjusted based on the results of HHS-RADV.

1.3.1 Overview of HHS-RADV Pilot Years Results (2015 and 2016 benefit year HHS-RADV)

During the 2015 benefit year HHS-RADV, issuers and IVA Entities experienced widespread challenges obtaining medical records. As such, HHS did not provide HHS-RADV results for the 2015 benefit year. However, based on feedback from stakeholders, HHS identified a number of process improvements and policy refinements that were incorporated in the 2016 benefit year HHS-RADV. For example, in the 2015 benefit year HHS-RADV pilot year, HHS required validation of demographic and enrollment (D&E) data for the full sample of 200 enrollees. Beginning with the 2016 benefit year HHS-RADV, HHS selected a subsample of 50 enrollees from the 200 enrollee sample for the IVA Entity to conduct D&E validation. This change was in response to IVA Entities encountering challenges validating D&E data on issuers' source systems and was intended to reduce the burden of this validation, as D&E errors identified in a subsample could still indicate a more systemic data submission issue for an issuer. HHS also provided the IVA sample to issuers six weeks earlier to allow more time for issuers to retrieve medical records.

For the 2016 benefit year HHS-RADV, 416 issuers participated[24] and were provided with illustrative HHS-RADV error rate results based on the current methodology. In our examination of the 2016 benefit year HHS-RADV results, HHS found that many issuers made significant improvements from the 2015 benefit year HHS-RADV results, but HHS's review of IVA submissions identified a number of serious concerns for some issuers with exceptionally high HCC group failure rates. Even though a large proportion of issuers passed the pairwise means tests, many issuers did not submit sufficient inpatient medical records, or submitted multiple

[21] The exception to the prospective application of HHS-RADV results is for exiting issuers, whose risk score error rates are applied to the calculated PLRS and risk adjustment transfer amounts for the benefit year being audited.
[22] See the Summary Report of 2017 Benefit Year HHS-RADV Adjustments to Risk Adjustment Transfers released on August 1, 2019 is available at: https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/BY2017-HHSRADV-Adjustments-to-RA-Transfers-Summary-Report.pdf.
[23] While the 2016 benefit year was a pilot year, issuers were provided illustrative 2016 benefit year HHS-RADV final results based on the application of the current methodology. The 2016 benefit year HHS-RADV results memo was made available to issuers in the HHS-RADV Audit Tool.
[24] HHS exempted from the 2016 benefit year HHS-RADV pilot small issuers with total premiums of $15 million or less and did not enforce participation in 2016 benefit year HHS-RADV for issuers that were not offering coverage in risk adjustment covered plans in the 2017 benefit year.

irrelevant medical records for each enrollee without providing a medical record that could substantiate the sampled enrollees' HCCs.

These findings resulted in very high HCC group failure rates for 77 issuers. Many of these issuers had HCC group failure rates that were outside of the modified confidence intervals for the 2016 benefit year HHS-RADV HCC groups and would have had adjustments to their respective risk scores had 2016 benefit year HHS-RADV been a non-pilot year. Because these issuers were contributing to the national metrics that created the confidence intervals, their results inappropriately inflated and skewed the national failure rate distributions, and would have impacted results for other issuers. For these reasons, issuers with exceptionally high HCC group failure rates (i.e., HCC group failure rates over 60 percent for the high HCC group, 50 percent for the medium HCC group, and 40 percent for low HCC group) were excluded from the national metrics for the 2016 benefit year HHS-RADV illustrative final results.[25] A summary of the modified 2016 benefit year HHS-RADV results with the exclusion of 77 issuers with exceptionally high failure rates is in Tables 1.1 and 1.2 below. By dropping these issuers from the national metrics for the 2016 benefit year, HHS increased the number of issuers that received non-zero error rates. The majority of the 77 dropped issuers were outliers and received positive error rates even though they were not counted as outliers in Table 1.2 (which do not include the 77 dropped issuers), and the majority of the 31 unique outlier issuers seen in Table 1.2 were outliers that received negative error rates. Due to the modifications to the final 2016 benefit year HHS-RADV results, the analyses documented in this paper primarily use the 2017 benefit year to test the policy considerations in this paper.

Table 1.1: 2016 Benefit Year HHS-RADV National Failure Rate Statistics

Number of Included HHS-RADV Issuers	Number of Issuers Dropped	Group	Mean	Standard Deviation	Lower Threshold	Upper Threshold
339	77	Low	0.142	0.109	-0.072	0.356
		Medium	0.251	0.114	0.028	0.475
		High	0.346	0.140	0.073	0.620

[25] While these issuers were dropped for purposes of calculating the national metrics for the 2016 benefit year HHS-RADV, CMS shared with these issuers their respective calculated error rates.

Table 1.2: 2016 Benefit Year HHS-RADV Number of HCC Groups Outliers at Issuer Level

Number of Included HHS-RADV Issuers	Number of Issuers Dropped	Group	Outliers Counts			
			Lower Bound	Upper Bound	Total	Unique Outliers[26]
339	77	Low	8	3	11	31
		Medium	6	4	10	
		High	14	0	14	
		Total	28	7	35	

1.3.2 Overview of First Non-Pilot Year of HHS-RADV Results (2017 Benefit Year HHS-RADV)

The 2017 benefit year was the first year that HHS operated the risk adjustment program in all 50 states and the District of Columbia. It also was the first non-pilot year of HHS-RADV such that HHS-RADV results were used to adjust risk scores and risk adjustment transfers.[27] All issuers of risk adjustment covered plans that did not have 500 or fewer billable member months or were not in liquidation were required to participate in 2017 benefit year HHS-RADV. A total of 595 out of 628 issuers of risk adjustment covered plans participated in 2017 benefit year HHS-RADV, an issuer participation rate of approximately 95 percent.[28] For the 2017 benefit year HHS-RADV, issuers substantially improved the retrieval and submission of adequate medical record documentation for validating HCCs compared to the 2016 benefit year HHS-RADV. However, in comparison to the 2016 benefit year HHS-RADV results, the rate of issuers who were identified as outliers increased for the 2017 benefit year HHS-RADV due to the changes in the distribution for each HCC grouping (see Table 1.4 below) and was described in the May 31, 2019 report.[29]

For 2017 benefit year HHS-RADV, the standard deviations from the mean failure rate for all three HCC failure rate groups were lower than the standard deviations for these failure rate groups in the 2016 benefit year HHS-RADV results, and fewer issuers were consistent outliers in multiple HCC groups in 2017 benefit year HHS-RADV. The 2017 benefit year HHS-RADV results showed shorter distances between HCC group failure rates and the mean group failure rate, and the magnitude of the adjustment factor in each HCC group and error rate also generally

[26] Since issuers can fail more than one HCC group, unique outliers refers to the number of issuers with at least one HCC group outlier.

[27] The one exception was for Massachusetts issuers, who were not able to participate in prior HHS-RADV pilot years because the state operated risk adjustment for those benefit years. Therefore, HHS made the 2017 benefit year HHS-RADV a pilot year for Massachusetts issuers. See the 2020 Payment Notice, 84 FR at 17508. While CMS provided illustrative 2017 benefit year HHS-RADV results to Massachusetts issuers, these results were not included in the national metrics and were not used to adjust risk scores or risk adjustment transfers.

[28] A total of 33 issuers of risk adjustment covered plans did not participate in 2017 benefit year HHS-RADV because they: (1) were exempt for having 500 or fewer billable member months statewide; (2) elected to receive a default data validation charge (DDVC); or (3) qualified for the liquidation exemption.

[29] https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/2017-Benefit-Year-HHS-Risk-Adjustment-Data-Validation-Results.pdf.

decreased in 2017 benefit year HHS-RADV compared to the 2016 benefit year HHS-RADV pilot. Thus, although there were more outliers in the 2017 benefit year HHS-RADV results, the error rates were lower in magnitude than those calculated during the 2016 benefit year HHS-RADV pilot, as expected.

The 2017 benefit year HHS-RADV results also included a number of issuers who exited all of the market risk pools in a state for the 2018 benefit year (exiting issuers). Eighty-one out of the 580 issuers[30] that participated and were used to calculate the national metrics for the 2017 benefit year HHS-RADV were exiting issuers. HHS-RADV results for the 81 exiting issuers were used to modify these issuers' 2017 benefit year risk scores and risk adjustment transfers for the applicable state market risk pools, rather than the 2018 benefit year risk scores and risk adjustment transfers.[31]

Table 1.3: 2017 Benefit Year National Failure Rate Statistics

Number of Included HHS-RADV Issuers	Number of MA Issuers Dropped	Group	Mean	Standard Deviation	Lower Threshold	Upper Threshold
580	15	Low	0.048	0.097	-0.143	0.238
		Medium	0.155	0.099	-0.040	0.349
		High	0.262	0.106	0.054	0.471

Table 1.4: 2017 Benefit Years HHS-RADV Number of HCC Groups Outliers at Issuer Level

Number of Included HHS-RADV Issuers	Number of MA Issuers Dropped	Group	Outliers Counts			
			Lower Bound	Upper Bound	Total	Unique Outliers
580	15	Low	15	34	49	110
		Medium	14	34	48	
		High	19	33	52	
		Total	48	101	149	

1.4 CONSIDERATION OF HHS-RADV CHANGES

In the following chapters of this white paper, we consider potential modifications to HHS-RADV based on our analysis of the above results, and comments received by stakeholders. Options described in these chapters were assessed independently of other potential policy changes being considered in this paper. For example, if we were to make the modifications to the

[30] Since the 2017 benefit year HHS-RADV was a pilot year for Massachusetts issuers, 15 Massachusetts issuers participated in 2017 benefit year HHS-RADV, but their HHS-RADV results were not used to set the national metrics.

[31] For the 2017 benefit year HHS-RADV, exiting issuers found to have a non-zero risk score error rate (i.e., that are identified as an outlier) will result in adjustments to 2017 benefit year risk scores and risk adjustment transfers. For the 2018 benefit year HHS-RADV and beyond, only those exiting issuers who are identified as having a positive risk score error rate outlier will result in adjustments to risk scores and risk adjustment transfers. See the 2020 Payment Notice, 84 FR at 17503.

outlier determination process contemplated in Chapter 3 of this paper, the determination of which issuers are outliers and the issuers' associated failure rates could be impacted. That determination may impact our policy approach with respect to the error rate adjustment options in Chapter 4 that are calculated on issuers' failure rates. Therefore, if we were to propose any of the options described in this paper in future rulemaking, we would reassess and re-evaluate the impact and trade-offs of the different options presented in this paper.

Because the analyses in this paper were primarily tested on one year of data (the 2017 benefit year HHS-RADV data), we note that further testing of future years of HHS-RADV data may change our perspective on some of the analysis in this paper. For example, many smaller issuers that were below the materiality threshold of less than $15 million in premiums for the benefit year were not required to participate in the 2016 benefit year HHS-RADV, but were all generally required to participate in 2017 benefit year HHS-RADV. These issuers' participation changed the population of issuers in the 2017 benefit year HHS-RADV results as compared to 2016 benefit year HHS-RADV results. For 2018 benefit year HHS-RADV and beyond, issuers within the materiality threshold will only be required to participate in HHS-RADV approximately every three years (barring any targeted audits). Therefore, in future benefit years, there could be fewer small issuers in the HHS-RADV results than in the 2017 benefit year HHS-RADV results. Likewise, in future benefit years of HHS-RADV results, changes to the risk adjustment models, changes to the population enrolled in risk adjustment covered plans, and changes in market participation may result in the identification of new trends or observations in future benefit years of HHS-RADV data that were not seen in the 2017 benefit year HHS-RADV data.[32] As future years of HHS-RADV data become available, we generally intend to continue to test the policy options described in this paper and identify areas for potential refinement and improvement in the HHS-RADV program.

[32] We also note that the benefit years used in the examples to illustrate the options being described in this white paper are only exemplary purposes.

2. HHS-RADV INITIAL VALIDATION AUDIT (IVA) SAMPLING

In this chapter, we review the background and purpose of HHS-RADV IVA sampling, our current sampling methodology, and feedback we have received on our current sampling methodology. We also discuss how we evaluate the HHS-RADV IVA sampling methodology by looking at precision and accuracy, and we outline several options for HHS-RADV IVA sample size refinement.

2.1 BACKGROUND AND PURPOSE OF HHS-RADV IVA SAMPLING

45 C.F.R. § 153.350(a) requires states, or HHS on behalf of states, to validate a statistically valid sample of risk adjustment data each year. Issuers' enrollee samples are the foundation of the HHS-RADV audit. These enrollee samples are also used to calculate an outlier issuer's error rate, which is applied to its risk scores and used to adjust risk adjustment transfers in the applicable state market risk pool. HHS sets the current enrollee sample sizes such that estimated risk score error rates will be statistically sound, enrollee-level risk score distributions will reflect enrollee characteristics for each issuer, and samples represent critical subpopulations of enrollees for each risk adjustment covered plan, such as enrollees with and without HCCs.

The 2015 Payment Notice stated that, after the initial years of HHS-RADV, HHS would evaluate our sampling assumptions using actual enrollee data to determine issuer-specific sample sizes.[33] In the 2020 Payment Notice, we proposed to vary the IVA sample size beginning with 2019 benefit year HHS-RADV based on each issuer's size, the prior year's HCC group failure rates, and sample precision.[34] However, at the time that we conducted analysis for the 2020 Payment Notice, we only had data from one pilot year of HHS-RADV and no data from small issuers because they were exempt from participating in the pilot years of HHS-RADV. In light of the limited available data and in response to stakeholder comments, we did not finalize any changes to our sampling methodology.[35]

2.2 FUTURE OF HHS-RADV IVA SAMPLING

HHS is contemplating several options to amend the methodology for enrollee sampling in future benefit years based on feedback and comments we have received from issuers and other HHS-RADV stakeholders. We have heard from some issuers that they want a larger sample size to improve precision, sample accuracy, and potentially decrease the impact of a single enrollee's results on their HCC group failure rates. Precision measures how close sample values are likely to be to each other. Accuracy measures how well the sample measurements match the true population value, without consideration of how close they are to each other.

At the same time, other issuers have asked for smaller sample sizes to reduce the administrative and financial burden associated with retrieving medical records and participating

[33] 75 FR at 13756-13759.
[34] 80 FR at 17492-17495.
[35] While we did not make changes to the sample size in the 2020 Payment Notice, we did finalize a change to our sampling approach to extend the application of the Neyman allocation to the 10th stratum. See 80 FR at 17492-17495.

in HHS-RADV. The next subsection in this chapter reviews the current HHS-RADV IVA sampling methodology and analyzes the precision and accuracy of 2017 benefit year HHS-RADV sample sizes. The following subsection describes the potential options being considered to adjust the current HHS-RADV sampling methodology.

2.3 CURRENT HHS-RADV IVA SAMPLING METHODOLOGY

2.3.1 Proxy Issuer Populations

HHS used two main data sources to design a sampling methodology and to estimate sample sizes for the 2015, 2016, 2017, and 2018 benefit years of HHS-RADV: MA-RADV net error rates and variance of net error; and Truven Health Analytics 2010 MarketScan® Commercial Claims and Encounters database-predicted expenditure data. HHS identified these sources as the most applicable empirical data that was available for the first years of the HHS-RADV program, because we did not have sufficient data from the HHS-operated risk adjustment program (i.e., enrollee-level EDGE data) at that time. HHS chose MA-RADV error rates because the MA-RADV program utilizes a similar HCC-based methodology to estimate risk of enrollees, and determines payment error rates based on evaluation of enrollee risk profiles and medical record validation. HHS determined that MarketScan® data was the best primary source that was available to approximate enrollee risk profiles in risk adjustment covered plans at the time, and used the MarketScan® data to calibrate the HHS-RA models.[36]

2.3.2 Stratification

In the individual market, the percent of enrollees with at least one HCC is approximately 22 percent – that is, approximately 78 percent of enrollees do not have an HCC.[37] Therefore, HHS determined that taking a simple random sample for HHS-RADV would not achieve the goal of evaluating higher risk enrollees within the population because a random sample would be composed primarily of enrollees with no HCCs or RXCs. Instead, using a simple age and risk score stratification, HHS divides each issuer's enrollee population into mutually exclusive groups or "strata" based on recorded risk scores, age, and presence of HCCs and RXCs, which are prescription drug categories that were added to HHS-RA adult models beginning with the 2018 benefit year. Statistical theory indicates that stratification of a population prior to sampling and the selection of more cases from strata with greater variance can increase the likelihood that the sample achieves targeted levels of confidence and precision relative to a simple random sample for which no stratification is performed. Based on the available data, HHS divides the relevant population into 10 strata, representing different age and risk score bands. This method of stratification is similar to that used in the MA-RADV program, which divides enrollees into three strata, representing low, medium, and high risk expenditures.

---

[36] HHS began incorporating enrollee-level HHS-RA data in its recalibration of the HHS-RA model beginning in the 2019 HHS-RA benefit year, as finalized in the 2019 Payment Notice. See 83 FR 16939-16941.

[37] See Figure 3 of https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/Summary-Report-Risk-Adjustment-2018.pdf.

Table 2.1 provides a listing of assigned strata by risk level for each age group. Strata 1-3 represent low, medium, and high-risk adults with the presence of at least one HCC or RXC. HHS updated the stratification logic for the three adult strata starting with the 2018 benefit year by adding the HCC or RXC condition.[38] RXCs are only used in the adult risk adjustment models and are not present or applicable for the remaining seven strata. Strata 4-6 represent low, medium, and high-risk children with the presence of at least one HCC. Strata 7-9 represent low, medium, and high-risk infants with the presence of at least one HCC. Stratum 10 consists of the No-HCC and No-RXC population and is not further stratified by age or risk level. Prior to 2019 benefit year HHS-RADV, strata 1-9 (enrollees with HCCs or RXCs) comprised two-thirds of issuers' 200 enrollee samples, with stratum 10 (enrollees without HCCs) comprising one-third of the sample. Beginning with the 2019 benefit HHS-RADV, the 10th stratum will no longer be constrained.[39]

Table 2.1: Stratification Mapping

HCC Stratum	Age	Risk Level	Stratum
1 or More HCC(s)	Adult	Low	1
		Medium	2
		High	3
	Child	Low	4
		Medium	5
		High	6
	Infant	Low	7
		Medium	8
		High	9
No HCCs	All	N/A	10

2.3.3 Target Precision and Confidence Interval

HHS targets a 10 percent relative sampling precision (or margin of error) for a two-sided 95 percent confidence interval. We established a 10 percent precision target based on a survey of

[38] HHS currently samples adults with RXCs or HCCs for strata 1 through 3. Because RXCs are not included in the calculation of HCC failures rates or error estimation, HHS is considering adjusting this sampling stratification methodology in future years.

[39] See 84 FR at 17494-17495.

guidance from the Office of Management and Budget (OMB), the Internal Revenue Service (IRS), and the HHS-developed Payment Error Rate Measurement (PERM) program.[40]

To meet the sampling precision target, each issuer needs to obtain a sample size such that 1.96[41] multiplied by the standard error, divided by their estimated adjusted risk score, equals 10 percent or less.

$$Precision = (1.96 * SE)/RS_{Adj}$$

In the formula above, *SE* is the standard error, which is the square root of the population variance, and RS_{Adj} is the estimated adjusted risk score. As sample size increases, standard error decreases, and precision improves (lower values of the precision measurement indicate a better precision) for a given estimated adjusted risk score.

2.3.4 Sample Size Calculation

HHS calculated the overall IVA sample size (n) using the following stratified mean estimator formula[42]:

$$n = \frac{(\sum_{h=1}^{H} N_h S_h)^2}{\sum_{h=1}^{H} N_h S_h^2 + \left(\frac{Prec \times Y}{CI}\right)^2}$$

- H is the number of strata;
- N_h is the population size of the h^{th} stratum;
- Y is the adjusted total risk score estimate, derived from MA-RADV data;
- S_h represents the standard deviation of risk score error amount for the h^{th} stratum, derived from MA-RADV data;
- *Prec* represents the desired precision level; and
- *CI* is the critical value for the confidence interval associated with the desired level, which is 1.96 for a two-sided 95 percent confidence interval.

Sample size precision analyses conducted using the formula above and proxy data from the MA-RADV program (Section 2.3.1) calculated a range of sample sizes to target 10 percent precision for a two-sided 95 percent confidence interval. Because there was no meaningful improvement in the estimated level of precision between a sample of 200 and larger sample sizes, HHS finalized a sample size of 200 enrollees for the IVA for issuers with enrollment equal to or greater than 4,000 enrollees.

To reduce financial and administrative burden for small issuers, HHS uses a Finite Population Correction (FPC) to calculate a smaller sample size for issuers with enrollment

---

[40] See https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicaid-and-CHIP-Compliance/PERM/index.html.

[41] Critical value for the two-sided 95 percent confidence level.

[42] The sample size formula can be found in Section 5.9: Cochran, William G., Sampling Techniques, third edition, John Wiley & Sons, 1977.

between 50 and 3,999. If an issuer has an enrollment of fewer than 50 enrollees, its sample size is equal to its enrollment. Issuers with 500 or fewer billable member months are exempt from HHS-RADV.[43] Additionally, beginning with the 2018 benefit year HHS-RADV, issuers that fall below the materiality threshold of $15 million in premiums will only have an IVA audit approximately once in three years (barring any risk-based triggers that warrant more frequent audits).[44]

The current enrollee sample size selected for the IVA is represented in the following Table 2.2.

Table 2.2: Current IVA Sample Sizes

Issuer Population Size (N)	IVA Sample Size (n)
$N \geq 4{,}000$	$n = 200$
$50 \leq N < 4{,}000$	$n = 200 \times Finite\ Population\ Correction\ (FPC)$ $FPC = (N - 200)/N$ If $(200 \times FPC) < 50, n = 50$
$N < 50$	$n = N$

2.3.5 Neyman Allocation

HHS calculates the individual sample size per stratum (n_h) using the Neyman optimal allocation method.[45] The Neyman method is designed to maximize precision, given a fixed sample size, using the Neyman formula:

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

- H is the number of strata,
- n is the total sample size (e.g., 200 for most issuers);
- N_h is the population size of the h^{th} stratum, and
- S_h represents the standard deviation of risk score error amount for the h^{th} stratum, derived from MA-RADV data.

The goal of sampling by strata is to pull samples that are not simply proportional to stratum size, as this may under-represent or over-represent the true drivers of risk score error. Instead, the Neyman formula determines the optimal number to be sampled from each stratum, proportional to each stratum's contribution to the total standard deviation of the population (i.e., larger samples are drawn from more variable strata). For the 2015, 2016, 2017, and 2018 benefit years of HHS-RADV, HHS only used the Neyman formula to calculate the sample size for strata 1-9, and set one-third of the sample size to be from the 10th stratum representing enrollees without

[43] 45 C.F.R. § 153.630(g)(1).
[44] 45 C.F.R. § 153.630(g)(2).
[45] See https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n324.xml.

HCCs.[46] Starting with the 2019 benefit year HHS-RADV, HHS will use the Neyman formula to determine the number of enrollees sampled in all 10 strata.[47]

2.3.6 Precision of Current Sample Sizes

HHS's goal is to achieve good precision and high accuracy of group failure rates because group failure rates determine whether an issuer is an outlier that will have its risk score adjusted to reflect its HHS-RADV error rate. HHS applies the risk score error rate to risk scores, which are used to adjust risk adjustment transfers.

Precision of the IVA sample is influenced by sample size, issuer population size, and risk score distribution. In the 2017 benefit year of HHS-RADV, most issuers reached the 10 percent group failure rate precision target. However, we found that issuers with sample sizes of fewer than 200 enrollees tended to have poorer precision than issuers with a sample size of 200 enrollees.

To forecast group failure rate precision for different sample sizes, we calculated the mean, standard deviation, and standard error of group failure rates from samples taken from the combined enrollee population of all 2017 benefit year HHS-RADV issuers (except for Massachusetts issuers). In this analysis, we define group failure rate precision as the half-width of the 95th percent confidence interval:

$$Precision = \frac{\left|CI_{Upper_{Bound}} - CI_{Lower_{Bound}}\right|}{2}$$

Figure 2.3 below shows the precision by HCC failure rate group for various sample sizes using two different methods: bootstrapping and independent sampling. Independent sampling requires drawing multiple samples without replacement from the issuer population, whereas bootstrapping involves taking one independent sample from the parent distribution and then drawing multiple, equal-sized samples with replacement from that initial sample. For each sample size, we calculated average group failure rates for the three HCC groups under both methods.

[46] See, e.g., 84 FR at 252.
[47] 80 FR at 17492-17495.

Figure 2.3: Precision Comparison: Bootstrap v. Independent Sampling Method

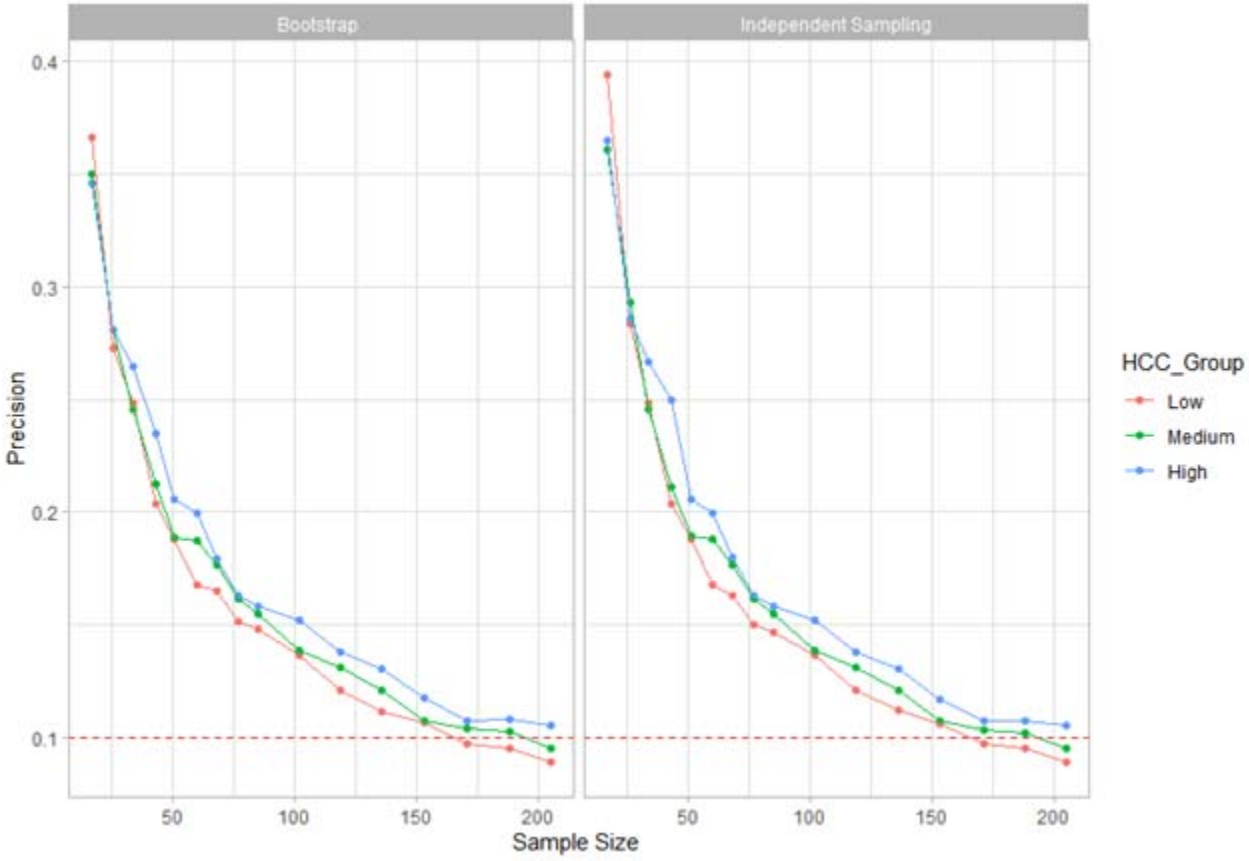
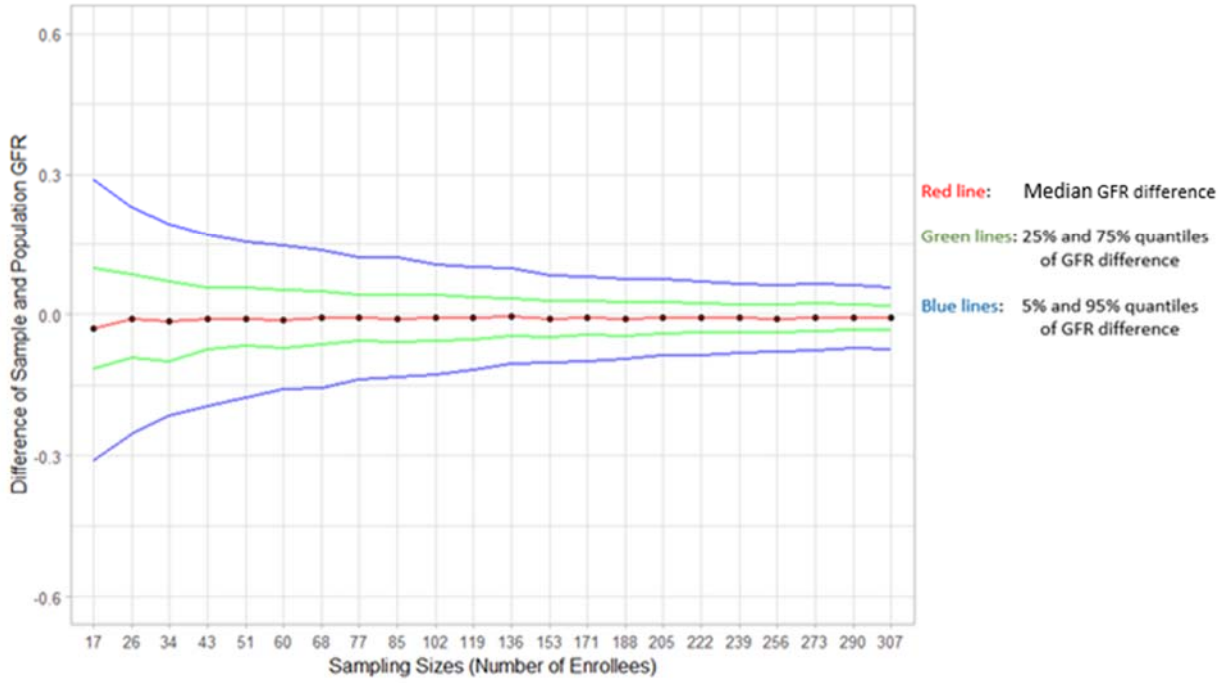


Figure 2.3 shows that precision improves (decreases in value) as sample size increases, and that on average, across all HCC groups, the current HHS-RADV sample size of 200 enrollees achieves the 10 percent precision target. We estimate that approximately 94 percent of issuers with a sample size of 200 enrollees meet the 10 percent precision target in at least one HCC group, and 60 percent of issuers with a sample size of 200 enrollees meet the target in all three HCC groups. For sample sizes greater than approximately 170 enrollees, the marginal improvement in precision is small.

2.3.7 Accuracy/Representativeness of Current Sample Sizes

In selecting HHS-RADV sample sizes, we also consider how well an issuer's IVA sample reflects their enrollee population, specifically, in the number and types of HCCs. Initial analysis using the combined enrollee population of all 2017 benefit year HHS-RADV issuers (except for Massachusetts issuers) indicates that the sample group failure rates accurately represent the group failure rates of the simulated issuer population. In Figure 2.4, we measure sample accuracy by the difference between the sample group failure rate and the issuer population group failure rate, across all three HCC failure rate groups. The range of differences between average sample and population group failure rates narrows and levels off around a sample size of 170 enrollees, and the gains in accuracy are small for larger sample sizes. There is more variability in accuracy for sample sizes below 170 enrollees, with standard errors increasing significantly for sample sizes below 50 enrollees.

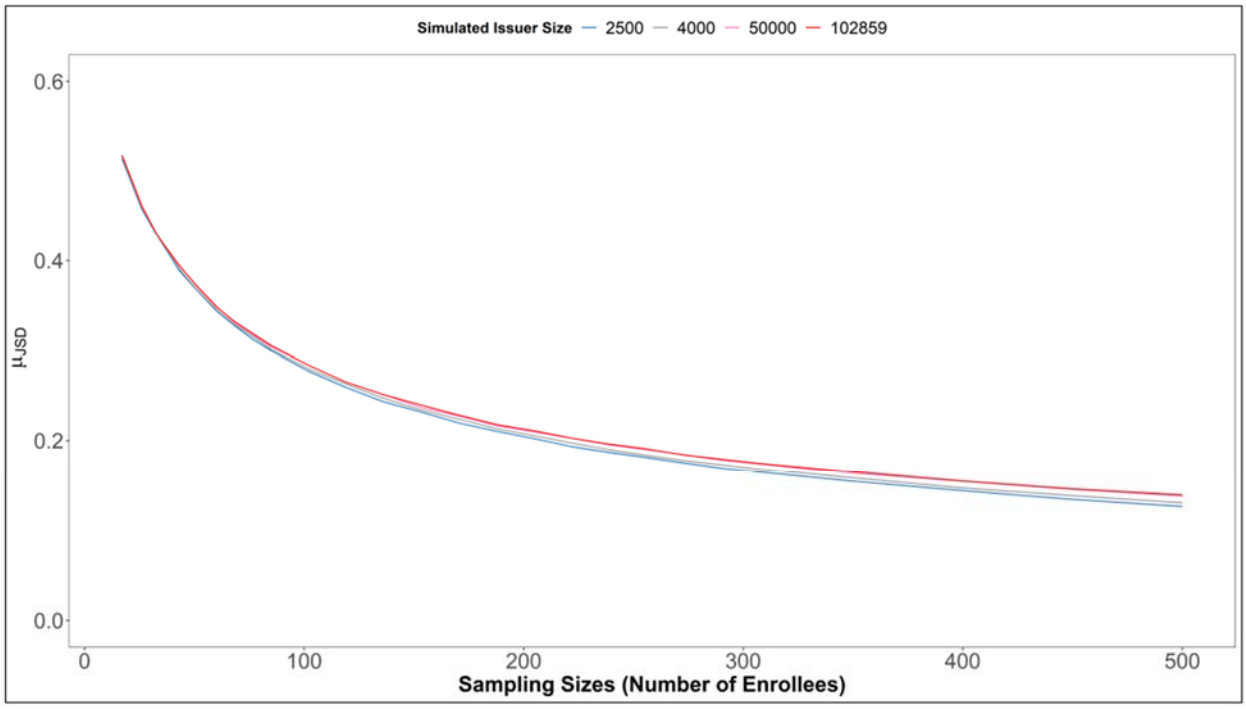
Figure 2.4: Sample Accuracy: Difference between Sample and Population Group Failure Rates (GFRs)



In Figure 2.5 below, we compared the probability of finding specific HCCs at different sample sizes to four different population sizes (2,500; 4,000; 50,000; and 105,577) simulated from the same combined population of 2017 benefit year HHS-RADV issuer enrollees (except for Massachusetts issuers). We used the Jensen-Shannon divergence (JSD)[48] metric to compare the probability distributions of the samples and the populations. As the value of the JSD decreases, the likelihood of finding the same HCCs in the simulated population and the sample taken from that population increases.

[48] Jianhua Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, 37(1):145–151, 1991.

Figure 2.5: Sample Accuracy: Difference between Sample and Population HCC Frequency Distribution



The discrepancies in the frequency at which specific HCCs occurred in the samples and simulated populations are inversely proportional to population size. However, for all simulated issuer sizes, we observe a substantial improvement in the degree to which the sample accurately represents the simulated population as sample size increases from roughly 25 to 100 enrollees. For samples larger than the current IVA sample size of 200 enrollees, there were only small marginal gains in the alignment of the sample and simulated population HCC frequency distributions. As such, our analysis shows that the current sample size of 200 enrollees achieves meaningful precision and accuracy, after which point there are diminishing improvements in these metrics and increased burden for issuers.

2.4 HHS-RADV IVA Sample Size Refinement

2.4.1 Goals for HHS-RADV IVA Sample Size Refinement

To refine our sampling methodology for future benefit years of HHS-RADV, we have the following goals:

- Ensure samples accurately represent issuer enrollee populations
- Increase the number of samples that meet the 10 percent precision target
- Minimize the administrative and financial burden on issuers, recognizing that any increase in sample size would increase the burden associated with retrieving and submitting relevant medical records, particularly for small issuers

Taking into consideration these competing goals, we recognize that any modification to sample sizes is unlikely to achieve all of them, and that some changes in sample sizes made to achieve some goals may counter others. For example, if issuer burden were not a concern,

increasing sample sizes for small issuers subject to the FPC under the current methodology could be a reasonable means of meaningfully improving sample precision for these issuers. However, an increase in sample size for issuer populations with low counts of enrollees with HCCs may not result in marked improvements, as we anticipate these issuers would generally have difficulty improving representativeness and precision. When considering modifications to the current sampling methodology, we aim to balance these competing goals.

2.4.2 Options for Sample Size Refinement

HHS is contemplating several options to amend the methodology for enrollee sampling in response to comments from issuers and other stakeholders. In response to some large issuers' requests for larger IVA sample sizes, HHS is considering allowing issuers to elect larger sample sizes despite evidence presented above that the current HHS-RADV IVA sample size of 200 enrollees is representative of underlying issuer populations and generally meets the 10 percent precision target. HHS cannot guarantee that a larger sample size will meaningfully improve the precision or representativeness of any issuer's sample. We previously proposed this option in the 2020 Payment Notice, but did not finalize this or any other changes to the IVA sample size in that rulemaking.[49] If this option is available in future benefit years, and an issuer elects a larger IVA sample size, we anticipate it would be limited by a maximum sample size to be determined by HHS, and the issuer would need to notify HHS of their chosen sample size by a date determined by HHS in advance of sample selection for the HHS-RADV benefit year. The number of enrollees sampled from strata 1-10 would still be calculated using the Neyman allocation method (Section 2.3.5) and the second validation audit (SVA) sample size would not increase in proportion to the elected IVA sample size – that is, the maximum SVA subsample would remain at 200.[50] We would consider the option for issuers to request larger sample sizes independent of, or alongside, one or more of the options described in Sections 2.4.2.1 through 2.4.2.3.

We are considering sample size refinements, which are outlined in Sections 2.4.2.1 through 2.4.2.3 below, that may help reduce operational burden for smaller issuers who do not fall within an exemption from HHS-RADV, while improving precision and representativeness of their IVA samples. In response to concerns from issuers about the administrative and financial burden of HHS-RADV, HHS currently uses three criteria to help identify small issuers for which the burden of sampling may be greater and the sample count of enrollees with HCCs may be too low to result in a representative sample:

---

[49] See 84 FR at 17492 to 17494. Also see 84 FR 227 at 252 to 256.

[50] The SVA sample sizes consist of an initial sample of 12 enrollees and expand, if necessary, to include 24, 50, and up to 100 in the event of failure of pairwise means testing. If an SVA sample size of 100 has poor precision, the sample may be expanded to the full IVA sample of 200. See Section 7.3.3 of the 2018 HHS-RADV Protocols at: https://www.regtap.info/reg_librarye.php?i=2904.

(1) Total annual premiums: Issuers at or below the $15 million premium materiality threshold only have an IVA approximately every three years (barring any risk-based triggers that warrant more frequent audits)[51]
(2) Enrollee population: Issuers with enrollee populations below 4,000 are subject to the FPC that reduces their sample size to between 200 and 50
(3) Billable member months: Issuers with 500 or fewer billable member months are exempt from HHS-RADV[52]

Most issuers that fall below the $15 million materiality threshold also have enrollee populations less than 4,000, but there are a few exceptions. Issuers with 500 or fewer billable member months typically have approximately 50 total enrollees.

We note that given application of the Neyman allocation to the 10th stratum beginning with the 2019 benefit year of HHS-RADV and the other potential policy changes presented in this paper, it is difficult to predict if sample size changes under these approaches will impact HHS-RADV failure rates, the determination of outlier status, and error rates.

2.4.2.1 Vary Sample Size Based on Issuers' Distance from the HCC Group Failure Rate Outlier Threshold and Precision

One option under consideration to adjust sampling would be to vary sample size based on issuers' distance from the HCC group failure rate outlier threshold and group failure rate precision using a prior year's HHS-RADV results. We previously proposed this method to adjust sampling in the 2020 Payment Notice, but did not finalize this or any other changes to sample size in that rulemaking.[53] Under this approach, HHS would increase the sample size for issuers that meet both of the following conditions:

(a) HCC group failure rates that fall outside 1.645 standard deviations of the mean in at least one HCC group;[54] and
(b) Group failure rate precision for the same HCC group above the 10 percent target.

Both conditions are evaluated using the HHS-RADV results for the benefit year two years prior to the benefit year for which the HHS-RADV sample is being drawn in at least one HCC group. Samples sizes for issuers who do not meet the above conditions would be determined using the current sampling methodology (described in 2.3.4).[55]

Issuers with HCC group failure rates that do not fall outside 1.645 standard deviations of the mean or that meet the 10 percent precision target in all HCC groups would still have a sample

[51] 84 FR at 17503.
[52] Although issuers exempt via the materiality threshold random sampling and with 500 or fewer billable member months statewide are exempt from performing an HHS-RADV initial validation audit, they are not exempt from transfer adjustments as a result of the application of HHS-RADV error rates in their state market risk pool.
[53] See 84 FR at 17492 to 17494. Also see 84 FR 227 at 252 to 256.
[54] 1.645 is the critical value for the two-sided 90 percent confidence level and σ is the standard deviation of the issuer population.
[55] As noted below, sample sizes for issuers who did not participate in HHS-RADV in the applicable prior year would also be calculated using the current sampling methodology.

size of 200, or smaller for issuers with enrollment between 50 and 3,999 enrollees, as the FPC would still apply. In the current error estimation methodology, we use a 95 percent confidence interval, or 1.96 standard deviations from the mean, to determine whether issuers are outliers in each HCC group, and ultimately to calculate error rates.[56] Expanding the confidence interval to 90 percent, or 1.645 standard deviations from the mean, to determine sample sizes would ensure that issuers that had higher- or lower-than-average HCC group failure rates in a prior year of HHS-RADV, but were not identified as group failure rate outliers due to poor precision in their samples, have larger sample sizes in future years of HHS-RADV. Due to the HHS-RADV timeline and the timing of the availability of the previous year's HHS-RADV results, this option would use HCC group failure rates from HHS-RADV results from the benefit year two years prior to the benefit year being audited to adjust the sample (e.g., 2018 benefit year results would determine 2020 benefit year HHS-RADV sampling).

Sample sizes for issuers that meet these conditions in at least one HCC group would be adjusted based on the distance of their current precision to the 10 percent target precision using the formula below:

$$n_{new} = n_{initial} * \left(\frac{Precision_{current}}{Precision_{new\_target}} \right)^2$$

Where $n_{initial}$ equals 400 for issuers with populations larger than 50,000 enrollees and $n_{initial}$ equals 200 for all other issuers. Extra-large issuers with poor precision and HCC group failure rates that fall outside 1.645 standard deviations of the mean would have larger sample size increases compared to medium-sized issuers. An issuer's final sample size would be the maximum n_{new} calculated for each of the HCC groups in which the issuer meets the group failure rate and precision criteria.

Issuers with $15 million or less in premiums who are selected to participate in HHS-RADV in a given benefit year could have much larger sample sizes under this methodology if they had poor precision in prior years of HHS-RADV. To limit the additional burden imposed on these issuers, we would use the approach that results in the smallest sample size from the sample size calculation methods below:

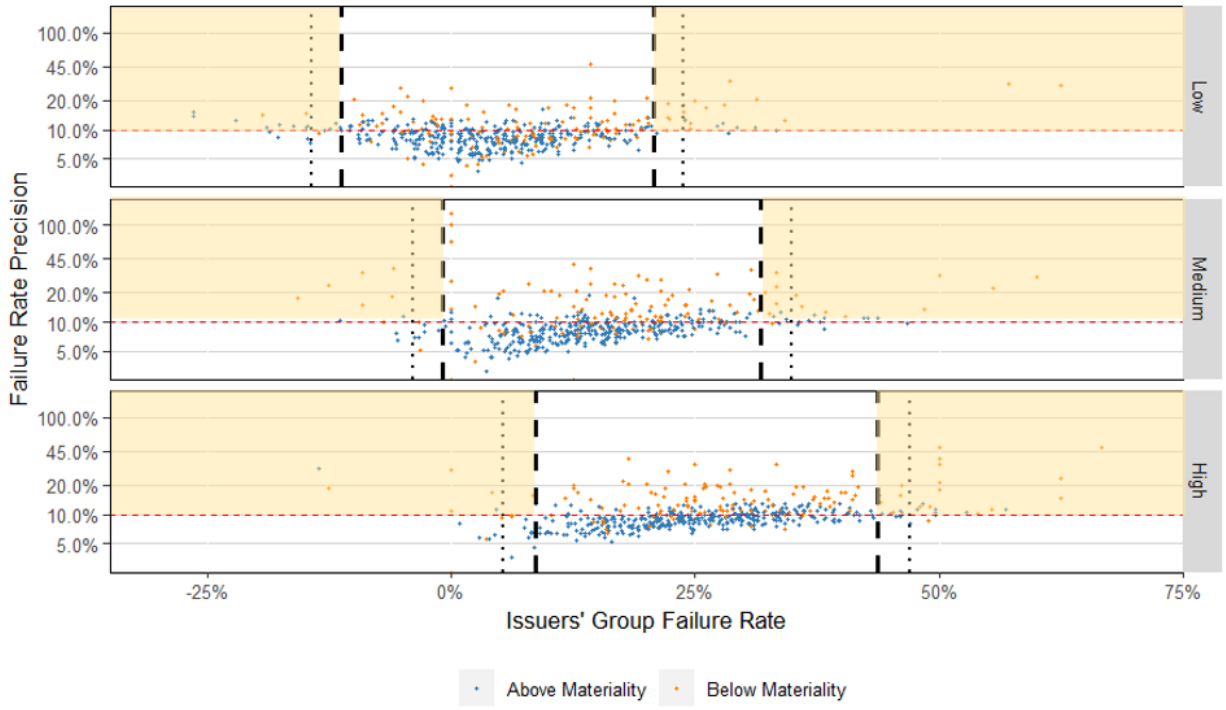
(1) The calculated sample size using the precision formula above; or
(2) The current sample size of 200 enrollees for issuers with enrollee population sizes greater than or equal to 4,000; or
(3) If an issuer has fewer than 200 enrollees, we would set their sample size equal to their population size in order to maximize precision.

We used the 2017 benefit year HHS-RADV results to test the option to vary sample size based on issuers' distance from the HCC group failure rate threshold and precision. We estimate that, out of the approximately 514 issuers expected to participate in HHS-RADV for benefit year

[56] As detailed above, the current sampling methodology targets a 10 percent relative precision (or margin of error) for a two-sided 95 percent confidence interval.

2020, approximately 92 issuers (57 of which we estimate would be issuers with $15 million or less in premiums, representing 38 percent of such issuers) would have their target sample size increased under this approach. Sample sizes for issuers that would experience sample size increases would range from approximately 117 to 462 enrollees.

Figure 2.6: Issuers Affected by Adjustment Based on Issuers' Distance from the HCC Group Failure Rate Outlier Threshold and Precision[57]



Issuers with HCC group failure rates that fall outside 1.645 standard deviations of the mean and with precision far from the 10 percent precision target (highlighted in Figure 2.6) would have an opportunity to improve their precision with the larger sample sizes under this option. Additionally, larger sample sizes could give issuers the opportunity to retrieve more accurate and complete medical records for HHS-RADV by capturing enrollees with HCCs that may have been missed in smaller samples.

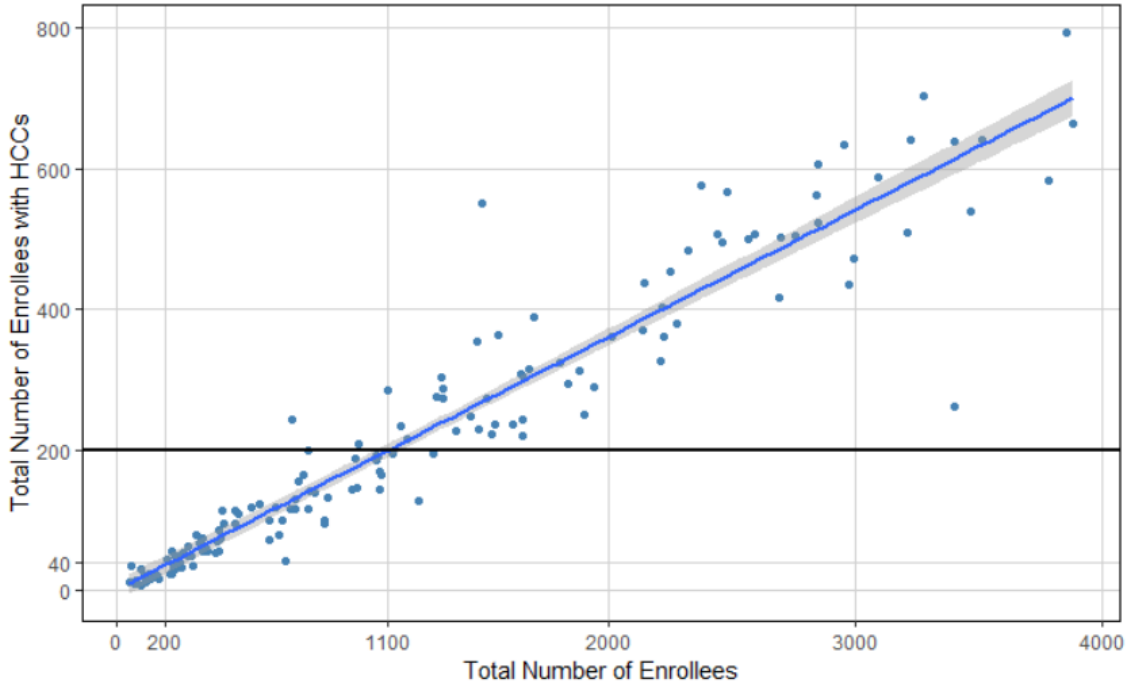
However, we have concerns about the potential burden associated with this option for small issuers with $15 million or less in premiums, poor precision, and HCC group failure rates that fall outside of the 90 percent confidence interval. These issuers would have larger sample sizes under this option for the benefit year(s) in which they are selected to participate in HHS-RADV. Financial and administrative burden could increase for those issuers when they may not have the

[57] Figure 2.6 shows failure rate precision results for 512 issuers offering risk adjustment covered plans in the 2018 benefit year with 2017 benefit year HHS-RADV results. Issuers "above materiality" have total annual premiums above $15 million. Issuers "below materiality" have total annual premiums at or below $15 million. Most issuers that fall below the $15 million materiality threshold also have enrollee populations less than 4,000, but there are a few exceptions. Beginning with 2018 benefit year HHS-RADV, issuers below the materiality threshold will be subject to random (or targeted) sampling. See 81 FR 94058 at 94104-94105.

capacity to retrieve more accurate medical records, and they may lack the additional enrollees needed to increase their sample size or meaningfully improve their precision.

Specifically, issuers with populations of fewer than 1,100 enrollees may not have enough enrollees with HCCs from which to sample. Based on an analysis of issuers in 2018 benefit year risk adjustment (Figure 2.7 below), approximately half of issuers with premiums at or below $15 million had fewer than 200 enrollees with HCCs. Since group failure rate precision is determined by the number of enrollees in the population with HCCs, a larger sample for these issuers would not necessarily improve their group failure rate precision. For example, increasing the sample size under this option from 150, as calculated under the FPC, to 200 for an issuer with a population between 50 and 3,999 enrollees may result in 50 more enrollees without HCCs being sampled, which would provide no meaningful improvement in group failure rate precision.

Figure 2.7: Total Number of Issuer Enrollees by Total number of Issuer Enrollees with HCCs



Moreover, this option requires using data from two years prior to adjust issuers' sample sizes, because sampling for each benefit year occurs before HCC group failure rates from the prior benefit year becomes available. For example, 2020 benefit year HHS-RADV sample sizes would need to be determined using 2018 benefit year HHS-RADV group failure rates because 2020 benefit year HHS-RADV sample sizes would need to be calculated before 2019 benefit year HCC group failure rate results are available. We recognize that another limitation of this approach is that using prior year HCC group failure rates may not be representative of an issuer's current population because population characteristics could change dramatically over two years, especially for small issuers. Additionally, this sample size adjustment would not be available for issuers that did not participate in HHS-RADV two years prior to when sample sizes are calculated because they would not have HCC group failure rate results available to calculate their

sample size. Sample sizes for these issuers would be calculated using the current sampling methodology (described in 2.3.4).

We are also considering whether we should evaluate combining HCC group failure rate data from multiple prior years of HHS-RADV once enough data becomes available, to potentially alleviate concerns that using one year of failure rate data may not be representative for determining sample sizes. However, this would not completely alleviate the concern of using prior year HCC group failure rates to represent issuers' current populations. Since we currently only have two years of HHS-RADV, only one of which is a non-pilot year, we believe more data and analysis of potential trends in failure rates across multiple years is needed. Additionally, because the HCCs in each HCC failure rate group, as well as the means and standard deviations of HCC group failure rates change year over year, the cutoff values for the 95 percent confidence interval in the current sampling methodology also vary year to year. This could make it difficult to combine issuers' historical failure rates across multiple years to determine which issuers to target for larger sample sizes under this option.

2.4.2.2 Re-evaluate the Standard Sample Size Using National Average HHS-RADV Error Rates Instead of Proxy Data from MA-RADV

To align our sampling methodology with risk adjustment program policy to use the most recently available program data as source data, under this approach, HHS would calculate sample size using national average HHS-RADV error rates instead of proxy data from MA-RADV (that we used to determine the current IVA sample size of 200). For issuers with population sizes of 4,000 enrollees or more, we would vary sample size based on issuer-specific population size, the distribution of enrollees between strata, and standard deviations of risk score errors among the 10 strata. The FPC would still be used to calculate a smaller sample size for issuers with enrollment between 50 and 3,999 enrollees. In addition, if an issuer has fewer than 50 enrollees, its sample size would remain equal to its enrollment.

Specifically, if an issuer's population size is 4,000 or more enrollees, then the same formula used to calculate the current IVA sample size from MA-RADV data would be used:

$$n = \frac{(\sum_{h=1}^{H} N_h S_h)^2}{\sum_{h=1}^{H} N_h S_h^2 + \left(\frac{Prec \times Y}{CI}\right)^2}$$

- N_h is the population size of the h^{th} stratum;
- Y is the adjusted total risk score estimate, that is, the adjusted total HHS-RADV risk score estimate using the average HHS-RADV error rate calculated by all issuers;
- S_h represents the standard deviation of risk score error amount for the h^{th} stratum.
- $Prec$ represents the desired precision level (still 10 percent); and
- CI is the confidence interval associated with the desired level, which is 1.96 for a two-sided 95 percent confidence level.

Under this option, an issuer's sample size would depend on its total population size and the distribution of enrollees and risk score errors between strata, so there would be no guarantee that its sample size would increase proportionally to its population size. Using the 2017 benefit year

HHS-RADV results, we estimate that approximately 330 issuers (all with populations of 4,000 or more enrollees) would have their sample sizes increased under this option out of the approximately 514 issuers expected to participate in HHS-RADV in benefit year 2020. We also estimate that approximately 31 issuers would have their sample size decreased, as this option would allow for customized sample sizes to achieve the targeted precision for each issuer. In total, we estimate this option would lead to an average sample size of 230 enrollees and an average sample size increase of 25 percent.

The data used to calculate the standard sample size under this option will better represent the population enrolled in risk adjustment covered plans than the MA-RADV data used to calculate the current standard sample size of 200. Further, though increasing sample sizes would increase operational burden for issuers, larger sample sizes could improve issuers' precision and help issuers obtain more accurate HHS-RADV results by capturing more enrollees with HCCs in the IVA sample.

Similar to the approach outlined in Section 2.4.2.1 that uses failure rates from two years prior, this option would also require using errors rates from two years prior, due to the timing of the calculation and release of HHS-RADV error rates and the timing of HHS-RADV sampling. However, this data is more recent and applicable to the enrollees in risk adjustment covered plans than the MA-RADV error rate data used in the current sampling approach. Additionally, unlike the option outlined in Section 2.4.2.1, this option uses the aggregated HHS-RADV results across all issuers, which may remediate some of the concerns introduced by using data from two years prior (e.g., use of prior year error rates may not be representative of an issuer's current population).

There are some other considerations for this option. If we were to determine the 2020 benefit year sample size based on results from the 2018 benefit year HHS-RADV data, the resulting sample size under this option could be smaller than what is forecast in this white paper using 2017 benefit year HHS-RADV data. For example, if the average error rate in 2018 benefit year HHS-RADV is significantly smaller than that of 2017 benefit year HHS-RADV, the resulting sample size(s) under this option would be smaller. Additionally, we currently have multiple benefit years of MA-RADV error rate data to use to predict sample sizes, but only have one non-pilot year of HHS-RADV data available to conduct this analysis and would only have two non-pilot years of HHS-RADV data if we implement this option for 2020 benefit year HHS-RADV sampling. In future years, once more HHS-RADV data becomes available, we would have more data to analyze potential trends in error rates across multiple years, and we could also further consider combining multiple benefit years of error rate data to calculate sample sizes. This paper does not outline options or offer an analysis related to the use of multiple benefit years of error rate data because there is currently only one non-pilot year of HHS-RADV data available. This option would also require the establishment of a different approach for determining samples sizes for issuers that did not participate in HHS-RADV two years prior to when sample sizes are calculated. Sample sizes for these issuers would be calculated using the current sampling methodology (described in Section 2.3.4).

2.4.2.3 Consider Other Sampling Options and Measures to Reduce Burden on Issuers with Small Populations

Another option to improve the precision and accuracy of samples for issuers with small populations is to maintain the current standard sample size of 200 enrollees for issuers who have sufficient enrollees in strata 1-10 to satisfy the Neyman allocation formula for that sample size. For issuers who do not have sufficient enrollees in strata 1-10 to satisfy the Neyman allocation formula, we would (a) determine an issuer-specific sample size that would reflect the sample size that satisfies the formula using their population total number of enrollees with HCCs or (b) consider adoption of additional criteria to exempt these issuers from HHS-RADV. Under this option, the FPC currently used to calculate the sample size for HIOS IDs with enrollment between 50 and 3,999 enrollees would no longer be used.

Figure 2.7 in Section 2.4.2.1 above (the first option for sample size refinement) indicates that issuers with populations of fewer than 1,100 enrollees may not have enough enrollees with HCCs from which to draw a sample to satisfy the Neyman allocation formula with a sample size of 200 enrollees. We chose a standard sample size of 200 enrollees based on our analysis described in Sections 2.3.6 and 2.3.7 above that increasing the sample size to more than 200 enrollees generally leads to minimal improvement in precision and accuracy. To determine which issuers would not be able to satisfy the Neyman allocation formula with a sample size of 200 enrollees under this option, for each issuer, we calculated sample sizes (n_h) for strata 1-9[58] (strata containing enrollees with HCCs) using the Neyman formula:

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

- H is the number of strata,
- n is the total sample size (set to 200 under this option),
- N_h is the population size of the h^{th} stratum, and
- S_h represents the standard deviation of risk score error amount for the h^{th} stratum.

Then, we determined which issuers had a total number of sampled enrollees in strata 1-9 greater than their total population of enrollees with HCCs (i.e., issuers that had a deficit of enrollees with HCCs from which to sample). Based on two analyses, one using MA-RADV error rate data (used to determine 2017 benefit year HHS-RADV samples) and another using 2017 benefit year HHS-RADV error rate data, we found that issuers with 1,100 or more enrollees or approximately 8,500 billable member months would have a sufficient total number of sampled enrollees in strata 1-9 to have an IVA stratified sample of 200 enrollees.

Under this option, issuers required to participate in HHS-RADV (that is, excluding issuers that meet the 500 or fewer billable member months exemption criterion) that we determine do not have enough enrollees with HCCs to satisfy the Neyman allocation formula for strata 1-10 with a sample size of 200 enrollees would have an issuer-specific sample size equal to the sum of

[58] We did not include stratum 10 in our analysis to determine which issuers would not be able to satisfy the Neyman allocation formula under this option because we assume that all issuers have sufficient enrollees without HCCs in their populations.

all of their enrollees with HCCs in each stratum 1-9 and the stratum 10 sample size that satisfies the Neyman allocation formula. This would give issuers with small populations who are required to participate in HHS-RADV an opportunity to improve their sample precision and accuracy. We anticipate that sample sizes would increase for some of these issuers and decrease for others when compared to the current sampling methodology (described in Section 2.3.4). Using the 2017 benefit year HHS-RADV results, we estimate the average sample size for these issuers would be approximately 86 enrollees. Each issuer unable to meet the required strata would have all enrollees with HCCs in their population sampled. We further note that we predict that most issuers that do not have enough enrollees with HCCs to satisfy the Neyman formula will likely fall under the $15 million materiality threshold exemption from HHS-RADV at 45 C.F.R. § 153.630(g)(2) and thus, would be subject to HHS-RADV approximately every three years (barring any risk-based triggers that would warrant more frequent audits).

Alternatively, we could consider adopting additional criteria to exempt these issuers from HHS-RADV, thereby reducing burden for issuers required to participate in HHS-RADV in circumstances where there is little or no potential to meaningfully increase group failure rate precision or improve representativeness of issuers' samples. For example, we could expand our current 500 billable member month exemption cutoff to provide relief for issuers with 8,500 or fewer billable member months. Billable member months, the current metric used for the HHS-RADV exemption at 45 C.F.R. § 153.620(g)(1), may more accurately represent plan enrollment than the count of enrollees, the metric used to identify issuers with low counts of enrollees with HCCs (see Section 2.4.2.1), and would align with the billable member month premium that we use to calculate risk adjustment transfers. We are interested in comments on the appropriateness of using billable member months as a metric for this new exemption cutoff in comparison to other metrics and the exemption cutoff value of 8,500 billable member months. Similar to issuers with 500 billable member months or fewer that are currently exempt under § 153.620(g)(1), issuers who qualify for this new exemption would not be exempt from the effects of HHS-RADV on transfer adjustments that may occur in their state market risk pool as a result of the application of HHS-RADV results. In addition, if we were to pursue this option and increase the number of issuers exempt from HHS-RADV, we would conduct targeted audits of exempt issuers under 45 C.F.R. § 153.620(c)[59] in order to mitigate the potential for gaming.

Although this alternative option would address the goal of decreasing burden for issuers below the new potential billable member month cutoff value[60], we have significant concerns about expanding the exemptions from HHS-RADV in this manner. Our main concern is the potential for gaming. In certain state market risk pools, some issuers below the new potential exemption cutoff may have a high risk score in comparison to the state market average risk score and HHS-RADV would not ensure those risk scores were not over-reported if this option were

[59] 45 C.F.R. § 153.620(c) states that HHS or its designee may audit an issuer of a risk adjustment covered plan to assess its compliance with the requirements of the risk adjustment program.

[60] Issuers that fall below the new potential billable member month exemption cutoff would also likely fall below the $15 million materiality threshold. However, under the new potential billable member month exemption, these issuers would not be required to participate in HHS-RADV approximately every three years.

adopted. This policy could also remove the incentives for these issuers to be vigilant in their coding practices and accurate in their EDGE data submissions. Further, as noted above, the existing materiality exemption at 45 C.F.R. § 153.630(g)(2) currently provides for decreased burden on issuers that would fall below the new potential billable member month exemption because they are currently only required to participate in HHS-RADV approximately every three years (barring any risk-based triggers that would warrant more frequent audits).[61]

Rather than look to adjust the sample size methodology, HHS is also considering different approaches to improve precision for issuers with low HCC counts, such as modifications to the outlier detection methodology described in Chapter 3 of this paper.

2.5 HHS's Perspective

HHS is interested in transitioning toward using HHS-RADV error rate data to replace MA-RADV proxy data and a preference for determining sample sizes in future years as outlined in Section 2.4.2.2. This would be consistent with HHS' risk adjustment program policy to use most recently available program data as source data, such as the transition in recent years from MarketScan® data to the most recently-available enrollee-level EDGE data for the annual calibration of the HHS risk adjustment models. We forecast that the average sample size calculated using HHS-RADV error rate data consistent with the approach in Section 2.4.2.2 for most issuers with populations of 4,000 or more enrollees would be relatively close in size to their samples of 200 under the current methodology. However, we only had one year of non-pilot HHS-RADV results available to forecast sample sizes under this option; future years of HHS-RADV may have smaller or larger error rates that may result in smaller or larger sample sizes for these issuers.

We acknowledge that the HHS-RADV operational timeline precludes our ability to make changes to the sampling methodology for the next applicable HHS-RADV benefit year (i.e., 2019 benefit year HHS-RADV), and that our analysis of policy options could benefit from the examination of several more years of HHS-RADV data that will become available before the start of 2020 benefit year HHS-RADV. While we previously requested comment in the 2020 Payment Notice on the possibility of permitting issuers to voluntarily increase sample sizes, we note that the current HHS-RADV sample size of 200 enrollees is representative of underlying issuer populations and generally meets the 10 percent precision target, as described in Sections 2.3.6 and 2.3.7. However, we continue to solicit feedback from issuers on whether electing a larger sample size than required by HHS is a desired approach. Lastly, we are interested in feedback from stakeholders on all of the options outlined in Section 2.4.2 that include: 1) varying issuers' samples for issuers with poor precision and who have an HCC group failure rate that falls outside 1.645 standard deviations of the mean; 2) utilizing HHS-RADV error rates in the calculation of issuer-specific sample sizes for issuers with 4,000 or more enrollees, while continuing use of the FPC for small issuers; or 3) considering other sampling options and measures (including potential expansion of HHS-RADV exemptions) to reduce burden on

[61] Ibid.

issuers with small populations. Under the third option, HHS would conduct targeted audits under 45 C.F.R. § 153.620(c) of issuers who are exempt from HHS-RADV to mitigate the potential for gaming that could result from expanding the exemptions from HHS-RADV.

3. MODIFICATIONS TO OUTLIER DETERMINATION

In this chapter, we review the process by which we determine whether an issuer qualifies as a failure rate outlier in HHS-RADV. This outlier determination process may prompt an adjustment to the issuer's risk score as calculated based on data reported on its EDGE server. We discuss two factors that may impact this process—HCC count and the interaction between HCC hierarchies and HCC failure rate groups—and explore several methodological changes that may help more precisely identify true outliers.

3.1 OVERVIEW OF FAILURE RATE OUTLIER DETERMINATION

As discussed in Section 1.2.1, the fourth step in the HHS-RADV process is error estimation. As a part of this stage, HHS determines the rate at which audit-validated HCCs[62] differ from EDGE-recorded HCCs and groups these HCCs into three (3) HCC failure rate groups (low, medium, and high). These rates are used first to establish a national standard, and then to determine whether individual issuers fall outside of an acceptable range of variation from that standard. Those issuers who fall outside of the acceptable range are termed outliers and their risk scores are adjusted based on the errors discovered during HHS-RADV.[63] The risk adjustment transfers for the applicable state market risk pool are modified in accordance with these risk score adjustments.[64] The specifics of this process are discussed below.

3.1.1 The Current Methodology

Under the current methodology, if an issuer's failure rate for an HCC group falls outside the confidence interval for the weighted mean failure rate for the HCC group, the issuer is considered an outlier for that HCC group. We use a 1.96 standard deviation cutoff, corresponding to a 95 percent confidence interval, to identify outliers. To calculate the thresholds for classifying an issuer's group failure rate as an outlier or not, the lower and upper limits of the confidence interval are computed as:

$$LB^G = \mu(GFR^G) - sigma\_cutoff * Sd(GFR^G)$$

$$UB^G = \mu(GFR^G) + sigma\_cutoff * Sd(GFR^G)$$

Where:

$\mu(GFR^G)$ and $Sd(GFR^G)$ are calculated as described in Section 1.2.3 of this paper.

$sigma\_cutoff$ is the parameter used to set the threshold for the outlier detection as the number of standard deviations away from the mean; in this case, 1.96.

---

[62] That is, HCCs validated by the IVA or SVA, as applicable.
[63] 45 C.F.R. § 153.350(b).
[64] 45 C.F.R. § 153.350(c).

LB^G, UB^G are the lower and upper thresholds to classify issuers as outliers or non-outliers for group *G*.

When an issuer's HCC group failure rate is an outlier, we reduce (or increase) the value of each of the applicable IVA sample enrollees' HCC coefficients by a proportion defined by the difference between the outlier issuer's failure rate for the HCC group and the national weighted mean failure rate for the HCC group. Formally, this adjustment amount is determined[65] by:

If $GFR_i^G > UB^G$ or $GFR_i^G < LB^G$:

Then $Flag_i^G = \text{"}outlier\text{"}$ and $Adjustment_i^G = GFR_i^G - \mu(GFR^G)$

If $GFR_i^G \leq UB^G$ and $GFR_i^G \geq LB^G$:

Then $Flag_i^G = \text{"}non\text{-}outlier\text{"}$ and $Adjustment_i^G = 0$

Where:

$Flag_i^G$ is the indicator if issuer *i*'s group failure rate for group *G* is located beyond a calculated threshold that we are using to classify issuers into "outliers" or "non-outliers" for group *G*.

$Adjustment_i^G$ is the calculated adjustment amount to adjust issuer *i*'s EDGE risk scores for all sampled HCCs in group *G*.

By this process, it is possible for an issuer to be flagged as an outlier and receive an adjustment in one of two ways. The issuer may be a positive outlier, meaning that the audit[66] was unable to validate a higher proportion of HCCs in a failure rate group than the national average; or the issuer may be a negative outlier.

The term "negative outliers" refers to issuers whose failure rate is demonstrated to be lower than the national average due to a failure rate lower than the lower threshold LB^G, indicating a statistically significant difference. Such outliers may occur if the audit resulted in a higher proportion of HCCs that are validated by the IVA in comparison to the national average, and if that difference is statistically significant. Negative outliers may also occur if the audit found a higher proportion of HCCs in the audit data that were not present in the EDGE data than the average issuer (i.e. "found HCCs"). If the number of found HCCs in a failure rate group exceeds the number of non-validated HCCs in that failure rate group for that issuer, it is possible for a negative failure rate to result.

---

[65] See 83 FR 16930 at 16963

[66] That is, the medical record retrieval and coding process performed by the IVA or SVA Entity, as applicable.

Found HCCs in an HCC grouping can happen for a variety of reasons.[67] At a high level, during the course of the medical record review by the IVA (or SVA as applicable), the IVA (or SVA) may find an HCC that is not associated with an HCC for an enrollee that was recorded in an issuer's EDGE server data. For example, a chronic condition may not have been diagnosed in the benefit year being audited, and therefore, the issuer may not have recorded that HCC in its EDGE server data. However, upon medical record review, that HCC may be found by the IVA (or SVA) and incorporated into an issuer's failure rate results in accordance with the guidelines on chronic, lifelong conditions outlined in the applicable benefit year's HHS-RADV Protocols.[68]

If an issuer is flagged as either a negative or positive outlier in a group, the adjustment value is applied to applicable HCCs for each enrollee on that issuer's EDGE server and the resulting HCC-level adjusted risk scores are summed for each enrollee to arrive at the enrollee adjustment, $Adjustment_{i,e}$, as described in Section 1.2.3 above. The enrollee-level adjustment is then aggregated for all of the issuer's enrollees on its EDGE server to arrive at the risk score error rate, which reflects the degree to which the risk score values found during the audit exceed or fall short of the risk score values reported through the EDGE server, relative to the national average rate at which EDGE and audit risk scores differ. HHS applies this value to the issuer's PLRS and adjusts the applicable benefit year's risk adjustment transfers for the state market risk pool(s) in question.

3.2 Addressing the Influence of HCC Count on Outlier Determination

Under the current methodology, we use national failure rate benchmarks to define a single set of confidence intervals that we apply to each of the three (3) HCC groups—based on the normal distribution—against which we validate all issuers' individual failure rates. Standard statistical theorems[69] state that as sample sizes increase, the sampling distribution of the means of those samples (in this case, the distribution of mean HCC group failure rates) will more closely approximate a normal distribution. At sufficient sample sizes, these theorems allow for normality to be assumed for statistical testing, ensuring the stability and reliability of results.

---

[67] Some stakeholders have suggested that including found HCCs in the calculation of failure rates may be counter to the goals of the HHS-RADV program. We disagree and believe that it is appropriate to include found HCCs to account to some extent for HCCs that were miscoded as another HCC within the same HCC hierarchy on EDGE. It is also necessary to ensure the HHS-operated risk adjustment program transfers funds from issuers with lower-than-average actuarial risk to issuers with higher-than-average actuarial risk. A further discussion of these types of miscoding scenarios is in Section 3.3 of this paper.

[68] See, e.g., Appendix E of the 2018 Benefit Year HHS-RADV Protocols, available at: https://www.regtap.info/reg_librarye.php?i=2904. As described in the 2018 Benefit Year HHS-RADV Protocols, CMS has implemented new HHS-RADV specific guidance related to chronic/lifelong conditions for 2018 HHS-RADV by updating the 2017 benefit year HHS-RADV 'Chronic Condition HCC' list with a simplified list of Lifelong Permanent Conditions, which is a subset of the conditions listed for 2017 HHS-RADV.

[69] In other words, the Central Limit Theorem (CLT). For background regarding the CLT, please see Ivo D. Dinov, Nicolas Christou, and Juana Sanchez. "Central limit theorem: New SOCR applet and demonstration activity." Journal of Statistics Education 16, no. 2 (2008). DOI: 10.1080/10691898.2008.11889560 .

As discussed in Chapter 2, we have already indirectly limited the inclusion of issuers below a certain number of enrollees through the exemption for issuers with 500 or fewer billable member months.[70] Although this exemption was primarily introduced to reduce disproportionate burden to issuers with fewer financial resources, it has the additional benefit of reducing the number of typical issuers who could be flagged as outliers due to low HCC counts in an HCC failure rate group. However, we believe that it is worthwhile to further mitigate the potential for typical issuers with low sample sizes to be flagged as outliers, on the basis that some samples may have too few HCCs to reliably determine whether their HCC failure rates in certain HCC groups are statistically different from the national means. As such, we are considering options to refine the outlier identification approach.

Although our sampling methodology is based on enrollee counts, the current error estimation methodology is based on HCC counts. For this reason, even though our analysis of sample size indicates that 200 enrollees provides sufficient precision on average as described in the previous chapter on sampling, the mismatch between the unit of analysis used for sampling and that used for error estimation may occasionally lead to fewer HCCs in an HCC group than may be necessary to reliably determine whether an issuer is statistically different from the national (average) HCC failure rate, as defined by static, national 95 percent confidence intervals. In effect, the national confidence interval may represent 95 percent confidence *in theory,* based on the assumptions that:

(1) all issuers come from a common population of issuers who are generally similar to each other regarding the obstacles they face in claim validation;

(2) the normal distribution is a fair approximation of the distribution of the failure rate; and

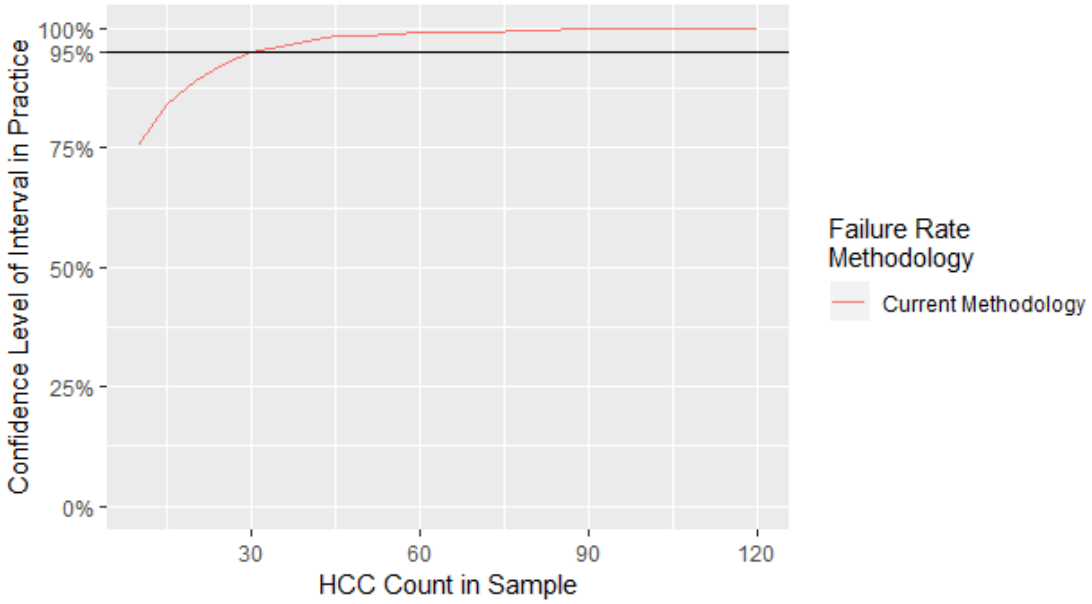
(3) issuers' samples are similar enough in HCC count that the precision of their failure rate estimates is about equal.

However, if any of these assumptions is substantially violated, the confidence level of the interval may diverge from 95 percent for some issuers in practice. For the purposes of this discussion, we will refer to the value that reflects the percentage confidence in practice as the *practical confidence level.*

To further examine this issue, we conducted an analysis in which we simulated the selection of samples from an average issuer using progressively smaller HCC counts (Figure 3.1). Through this process, we identified a threshold of 30 HCCs in an HCC group reported in EDGE data for a sample of enrollees as the threshold where the practical confidence level of the national confidence interval was lower than the theoretical 95 percent. This analysis indicates that the current methodology may be overly sensitive for issuers with fewer than 30 HCCs in an HCC group.

[70] 45 C.F.R. § 153.630(g)(1).

Figure 3.1. Resampling Simulation Comparing Theoretical and Practical Confidence Level of Current HHS-RADV Results by the Number of HCCs in a Sample



The result of this analysis suggests that issuer-specific HCC counts within an HCC group can help refine the methodology to more precisely identify true outliers. Using a single, static confidence interval across all issuers may have the impact that some issuers with low HCC counts within an HCC group could be identified as outliers, although at the population-level (rather than sample-level), they may be a typical issuer, that is, an issuer with a population-level failure rate indistinguishable from the national average. In effect, the static national 95 percent confidence interval may be too narrow to determine statistical significance at the desired confidence level when HCC count is below 30.

Furthermore, given that the national confidence intervals are static and do not vary based on individual issuer sample characteristics, there is the potential for atypical issuers with population-level failure rates that are very far from the national mean to have sample failure rates that fall within the confidence interval. In such cases, these issuers would not be identified as outliers in HHS-RADV.

Either of these situations will have an impact on other issuers in the state market risk pool. In the first case, some typical issuers, because of low HCC counts, may be identified as negative outliers under the current methodology, prompting an increase in their risk score, a higher payment or lower charge for the outlier issuer, and therefore lower payments or higher charges for other issuers in their state market risk pool. If these issuers are instead identified as positive outliers under the current methodology, their risk score will be decreased, resulting in lower payments or higher charges for the outlier issuer and increased payments or decreased charges to other issuers in their state market risk pool. Other issuers in the state market risk pool would also be impacted if atypical issuers whose population-level failure rates are above the national mean are not identified as positive outliers due to the static nature of the national confidence intervals, failing to prompt adjustments to transfers. Atypical issuers with failure rates well below the national mean could be harmed if they were not identified as negative outliers due to static

national confidence intervals that do not vary based on HCC count and other sample characteristics.

As the single set of static national confidence intervals appears to yield intervals that are too narrow for some issuers' HCC counts and too wide for other issuers' HCC counts, we believe that a methodology that would scale confidence intervals across the full range of HCC counts in our issuer population would permit more precise identification of true outliers. To this end, this chapter explores several alternatives to modify the current error estimation methodology.

3.2.1 Basic Modifications to Current Methodology Considered

The alternative methodologies described in this section reflect only minor changes to the current error estimation process. Although these methodologies vary in how well they improve the identification of true outliers, they share the benefit of maintaining a fair amount of the current error estimation methodology and potentially reducing any confusion and uncertainty generated by the adoption of a completely new methodology. The first method would establish multiple sets of national confidence intervals to account for issuers with varying numbers of HCCs in a grouping, and the second method would create issuer-specific bootstrapped confidence intervals.

3.2.1.1 Establish Multiple Sets of National Confidence Intervals

We explored creating multiple sets of national confidence intervals based on the number of HCCs present in an HCC failure rate group. Under this option, we would calculate two sets of national benchmarks for HHS-RADV by subdividing the population of issuers by the number of HCCs present in each issuer's failure rate groups: one for the category of issuers with high HCC counts, and one for the category of issuers with low HCC counts. We would then assess each category of issuer and HCC group based on the relevant confidence interval applicable to the category. Preliminary analysis suggests that, due to the natural increase in the size of the standard deviation of sample means when sample HCC counts are smaller, the low HCC count confidence intervals would be wider than the high HCC count confidence intervals, leading to fewer low HCC count issuers being identified as outliers compared to our current methodology. For example, simulations on the 2017 benefit year HHS-RADV data produced the following confidence interval limits for the high-failure rate HCC group:

Table 3.2. National Benchmarks for 2017 HHS-RADV Data under the Current Methodology and the Multiple Confidence Interval (MCI) Methodology

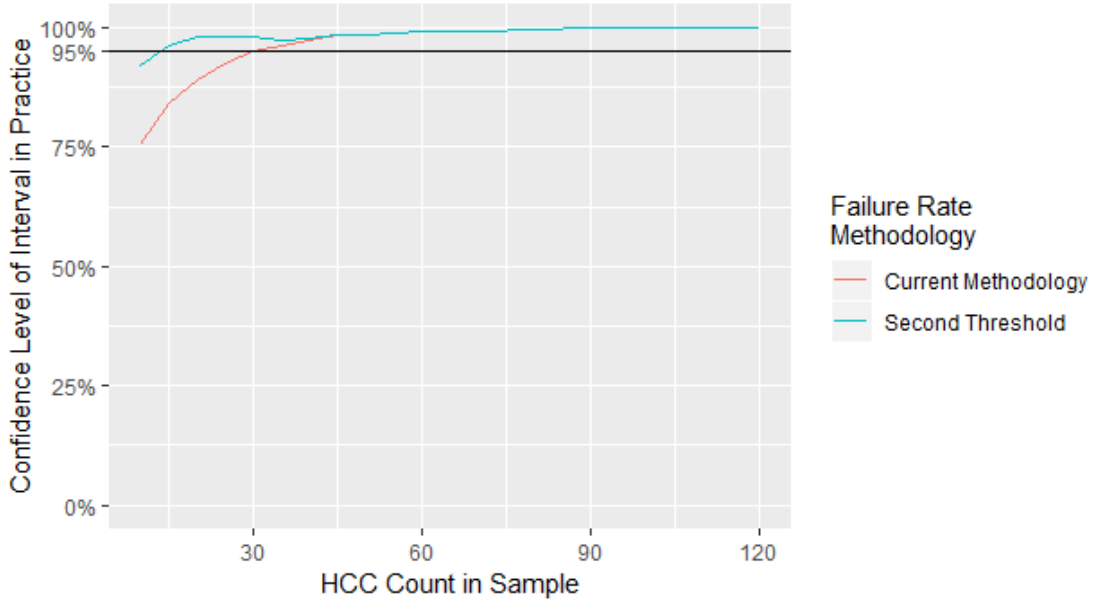
Issuer HCC Count Group	HCC Failure Rate Group	Group Mean Failure Rate	Standard Deviation	Confidence Interval Bounds	
				Lower	Upper
Current Method	Low	0.048	0.097	-0.143	0.238
	Medium	0.155	0.099	-0.040	0.349
	High	0.262	0.106	0.054	0.471
MCI Method for High HCC Counts (≥30)	Low	0.047	0.096	-0.142	0.235
	Medium	0.155	0.097	-0.036	0.345
	High	0.262	0.104	0.058	0.466
MCI Method for Low HCC Counts(<30)	Low	0.117	0.145	-0.167	0.401
	Medium	0.157	0.176	-0.188	0.503
	High	0.279	0.184	-0.083	0.640

In our simulated analysis of this method, the standard deviations were larger for the group of issuers with HCC counts that were less than 30 and, consequently, the confidence intervals were wider, with lower values for their lower bounds and higher values for their upper bounds. The confidence intervals and standard deviations for the group of issuers with 30 or more HCCs in an HCC group were about the same as the values under the current methodology.

The increased range for low HCC count issuers demonstrated in Table 3.2 reflects a national standard that allows for a greater degree of variability when HCC counts are low. As such, fewer issuers to whom these wider confidence intervals are applied will be flagged as outliers, and more of these issuers will have error rate values of zero, likely reducing the total absolute value of HHS-RADV transfer adjustments within state market risk pools, albeit only slightly.

As compared to the current methodology, the development of this second set of national benchmarks would allow the practical confidence level for samples with fewer than 30 HCCs to better approximate the theoretical 95 percent confidence level, as demonstrated by Figure 3.3.

Figure 3.3. Resampling Simulation Comparing Theoretical and Practical Confidence Levels of Current and Minimum HCC Count HHS-RADV Methodology by the Number of HCCs in a Sample



Establishing multiple sets of confidence intervals based on subsets of issuers appears to reduce the rate at which issuers with low HCC counts may be flagged as outliers. However, this option does not directly scale the width of the confidence interval according to the HCC count of the issuer, which we believe would be more likely to improve our ability to identify true outliers. Based on the analysis we have conducted thus far, we believe that is a significant shortcoming of this option.

3.2.1.2 Issuer-Specific Bootstrapped Confidence Intervals

In our search for a methodology that would directly scale the width of the confidence interval according to the HCC count of the issuer, we explored bootstrapping—a technique that avoids any assumptions regarding the underlying distribution of the failure rate metric. Bootstrapping is

a resampling simulation methodology that uses observed data—as opposed to formulas based on the central limit theorem—to provide information regarding the level of confidence we can have in an estimated value.[71]

Under this option, we would no longer calculate a single set of confidence intervals around the national mean and compare each issuer's failure rate to that confidence interval. Instead, HHS would calculate confidence intervals around each issuer's failure rate estimated for each HCC group, reflecting the stability of the estimate of that issuer's failure rate based on the issuer's data and HCC count. If, after bootstrapping, an issuer's confidence intervals do not include the national mean failure rate for any of the confidence intervals' respective HCC group, we would be able to conclude that the issuer's failure rate for that HCC group was significantly different from the national average, and the issuer's failure rate would be considered an outlier for that HCC group.

The process for bootstrapping individual issuer confidence intervals begins with the data for one issuer's HHS-RADV sample. For example, from the sample of 200 enrollees for that one issuer, we would draw a simulation sample with replacement equal in size to the original sample. Because this simulation sample is drawn with replacement, it will contain instances where, by random chance, a particular enrollee is included in the simulation sample more than once. For example, if an original HHS-RADV sample contained enrollees A, B, and C, the following samples would all be valid simulation samples: A-A-A; B-B-B; C-C-C; A-A-B; A-A-C; A-B-C; B-B-A; and so on.

Once we have taken a single simulation sample, the failure rates for that sample would be calculated and logged. We would then repeat the resampling and failure rate calculation process 1,000 times, resulting in a record of failure rates for 1,000 resamples of the original sample. Then, within each failure rate HCC group (high, medium, and low), we would put all of the failure rates in order by size and find cutoffs for the middle 95 percent of resampled failure rates. The cutoffs would serve as the upper and lower bounds of the confidence interval. This process would be repeated for every issuer's sample, arriving at issuer-specific confidence intervals for each HCC group. Due to the resampling procedure, the range of simulation sample means for issuers with lower HCC counts ought to be greater than the range of simulation sample means for issuers with higher HCC counts, because even a single randomly sampled validation failure for a low HCC count sample will have a greater impact on the estimate of the failure rate for that sample than a single validation failure for a high HCC count sample. For this reason, issuers with fewer HCCs in their samples ought to have wider confidence intervals than issuers with more HCCs in their samples, providing greater allowed variation for low HCC count issuers, while performing the same outlier determination process.

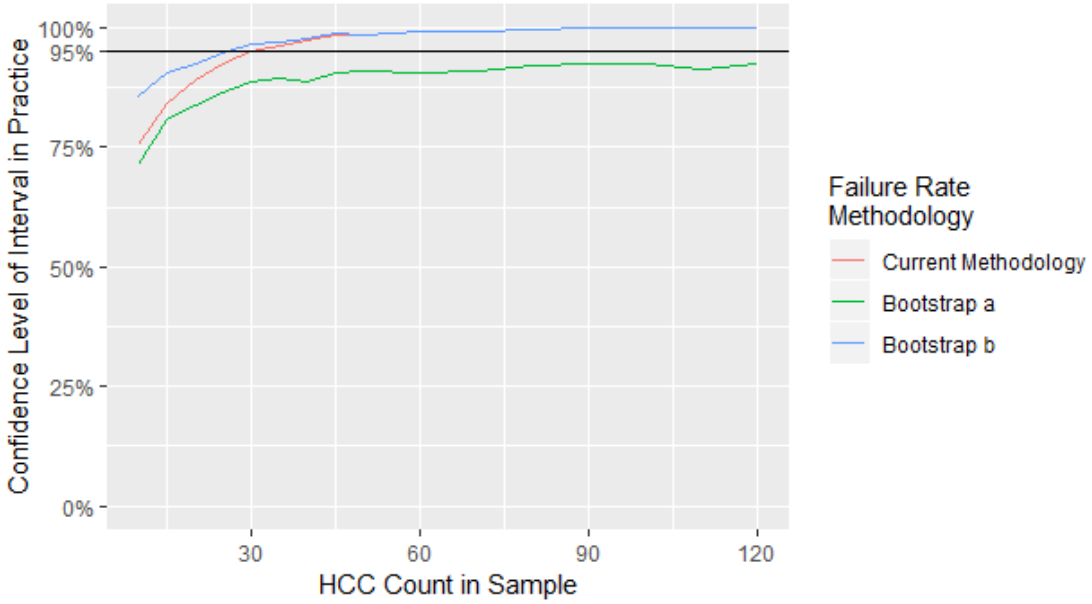
We explored two possible implementations of this process. In bootstrapping method *a,* we would apply this process to all issuers, regardless of HCC count. In bootstrapping method *b,* we would still apply this process to issuers regardless of HCC count, but only if the issuer were

---

[71] Phillip Good, *Introduction to Statistics Through Resampling Methods and R/S-Plus* (Hoboken, NJ: John Wiley & Sons, 2005).

initially flagged as an outlier by the current methodology. This second approach would essentially amount to using the bootstrapping estimation methodology to double-check the current methodology and ensure that a particular outlier identification was robust. However, our simulation of these two bootstrapping methods found that bootstrapping method *a* resulted in more cases of typical issuers who are identified as outliers than the current method, while bootstrapping method *b* improved upon the current method by only a small amount (Figure 3.4).

Figure 3.4. Resampling Simulation Comparing Theoretical and Practical Confidence Levels of Current and Bootstrapping HHS-RADV Methodology by the Number of HCCs in a Sample



Therefore, HHS does not believe that a bootstrapped resampling approach is appropriate to address the low HCC count issue. Furthermore, the calculation and presentation of the confidence interval thresholds using the bootstrapping method would not be based on formulas, and that lack of transparency could make it difficult for issuers to predict and incorporate HHS-RADV outcomes into rate setting assumptions.

3.2.2 Alternative Methodologies Based on Classical Statistics Considered

In our effort to explore longer-term options that provide a holistic solution to the low HCC count issue (subjecting issuers to a common outlier identification process, reduce the rate at which typical issuers could be flagged as outliers, and increase our ability to detect *atypical* issuers as true outliers), we examined two statistical options that would allow us to adjust for the HCC count at each issuer formulaically.

To accomplish this, we first decomposed our current measure of failure rate into its constituent parts to examine other ways in which mismatches between EDGE and audit data might be tested. In this vein, all coding scenarios between EDGE and audit results for each HCC taken separately can be represented by the following contingency table (Table 3.5). In this table, we have two sets of codings of the same data where the coding is dichotomous, i.e. either "present" or "absent".

Table 3.5. Cross-Tabulation of Possible Coding Scenarios for HCCs in EDGE and Audit Data

		Audit Data		
		Absent	Present	Total
EDGE	Absent	$absentHCC$	$newFoundHCC_{IVA}$	$F - freq_{EDGE}$
	Present	$missingHCC_{IVA}$	$validatedHCC$	$freq_{EDGE}$
	Total	$F - freq_{IVA}$	$freq_{IVA}$	F

In this contingency table,

- $F = n * k$, where
  - n is the number of enrollees in the IVA sample for the issuer;
  - k is the number of distinct HCCs under consideration, e.g. k is equal to 1 if each of the 127 HCCs evaluated in HHS-RADV is tested individually, or is equal to how ever many HCCs are in the low, medium, or high failure rate group, if HCCs are grouped before evaluation, as in the current methodology;
- $freq_{EDGE}$ and $freq_{IVA}$ follow the same definitions as in the current methodology: the number of occurrences of that HCC (or HCCs in an HCC group, if grouping is used) among sampled enrollees in EDGE and audit data, respectively;
- $absentHCC$ is k times the number of enrollees without that HCC (or HCCs in an HCC group) in *both* EDGE and audit data;
- $newFoundHCC_{IVA}$ is the number of occurrences of that HCC (or HCCs in an HCC group) that were identified during IVA or SVA, but were not present in the original EDGE data among sampled enrollees;
- $missingHCC_{IVA}$ is the number of occurrences of that HCC (or HCCs in an HCC group) that were present in the original EDGE data, but were not validated in audit data among sampled enrollees; and
- $validatedHCC$ is the number of occurrences of that HCC (or HCCs in an HCC group) that were present in the original EDGE data and were validated in audit data among sampled enrollees.

As discussed in the 2018 benefit year HHS-RADV protocols,[72] our current failure rate metric is calculated as:

[72] 2018 Benefit Year Protocols: PPACA HHS Risk Adjustment Data Validation, Version 7.0 (June 24, 2019), available at https://www.regtap.info/reg_librarye.php?i=2904.

$$FR = 1 - \frac{freq_{IVA}}{freq_{EDGE}}$$

Or, through algebraic operations:

$$FR = \frac{missingHCC_{IVA} - newFoundHCC_{IVA}}{freq_{EDGE}}$$

Our ability to separate *absentHCC*, *newFoundHCC*$_{IVA}$, *missingHCC*$_{IVA}$, and *validatedHCC* opens up the possibility of additional statistical techniques beyond our current methodology, and would allow us to make more substantial and targeted changes to refine the process of detecting outliers.

3.2.2.1 Binomial Distribution Methodology

Under this option, we would no longer assess issuers based on their failure rates. Instead, we would independently examine whether (1) HCCs in EDGE were validated in audit data, and (2) HCCs in the audit data were newly found HCCs. Because we would no longer use failure rates as a means of determining risk score error rates if we were to adopt the Binomial Distribution methodology, we would need to develop of a new methodology to adjust risk scores and risk adjustment transfers to reflect HHS-RADV results.

Because there are three HHS-RADV outcomes represented by the failure rate metric: $newFoundHCC_{IVA}$, $missingHCC_{IVA}$, and $validatedHCC$, determining *a priori* how three related outcomes impact the distribution of a metric is very difficult statistically. To address this challenge under the current methodology, we assume that the distribution of the failure rate approaches a normal distribution for large enough sample sizes, allowing us to apply confidence intervals based on this distribution across all issuers.

Although we must make assumptions regarding the shape of the current methodology's failure rate sampling distribution, it is easier to determine exactly how metrics will be distributed when only two—rather than three—outcomes are considered. In these cases, the metric describing the two outcomes would be distributed according to the binomial distribution. For example, we could calculate a binomially-distributed "non-validation rate" (*NVR*) as:

$$NVR = \frac{missingHCC_{IVA}}{missingHCC_{IVA} + validatedHCC} = \frac{missingHCC_{IVA}}{freq_{EDGE}}$$

We know that this *NVR* will be binomially distributed because the validation of any given HCC recorded in EDGE ought to be statistically independent from the validation of any other given HCC, and only two outcomes— $missingHCC_{IVA}$ and $validatedHCC$—influence this metric.

In the same way, we could calculate a binomially-distributed "new found rate" (*FndR*) as:

$$FndR = \frac{newFoundHCC_{IVA}}{newFoundHCC_{IVA} + validatedHCC} = \frac{newFoundHCC_{IVA}}{freq_{IVA}}$$

Both of these metrics could be assessed for outlier status based on confidence intervals around issuers' estimates using the binomial distribution, rather than the normal distribution.

Furthermore, because this distribution is defined by sample size and the magnitude of the metric (i.e., NVR and $FndR$), rather than the national standard deviation, the widths of these confidence intervals would vary for each issuer and HCC group based on how extreme the estimates of the non-validation and new-found rates are, and how large the HCC count is for an HCC group. As such, smaller HCC counts would receive wider confidence intervals and larger HCC counts would receive narrower confidence intervals.

The lower limit of the confidence interval for each issuer for each HCC group would be calculated as:

$$LL = \frac{2dp + z^2 - \left(z\sqrt{z^2 - \frac{1}{d} + 4dp(1-p) + (4p-2)} + 1\right)}{2(d + z^2)}$$

And the upper limit as:

$$UL = \frac{2dp + z^2 + \left(z\sqrt{z^2 - \frac{1}{d} + 4dp(1-p) - (4p-2)} + 1\right)}{2(d + z^2)}$$

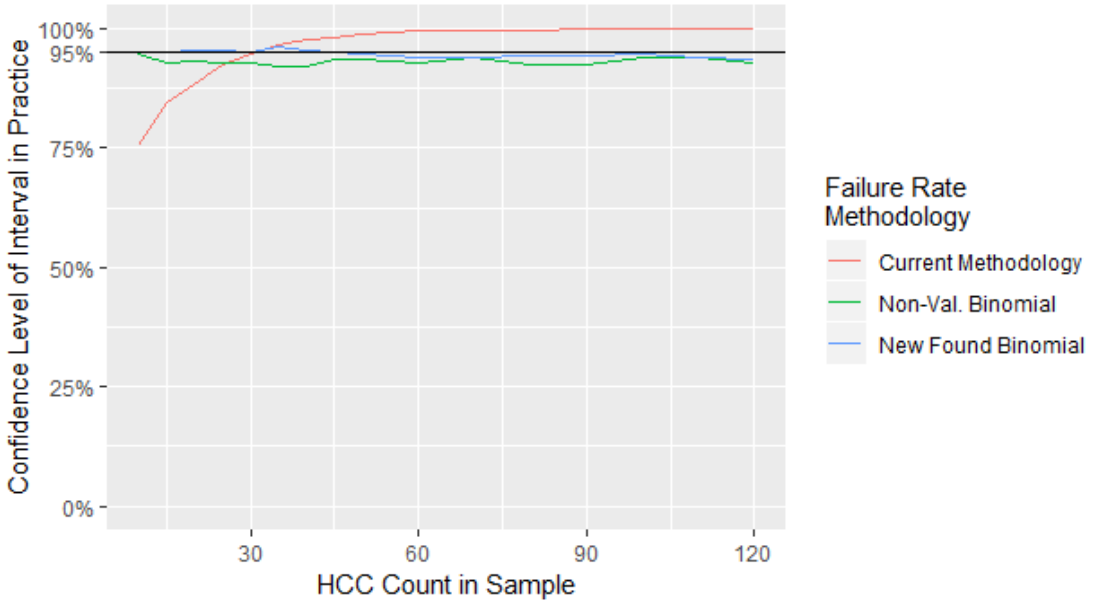
Where:

- d is the denominator of *NVR* or *FndR,* that is, $freq_{EDGE}$ and $freq_{IVA}$, respectively.
- p is the *NVR* or *FndR,* whichever is under consideration.
- z is the z-value cutoff for a 95 percent confidence interval: 1.96

As with the bootstrapping methodology, these confidence intervals are around the sample estimates, rather than around the national mean. If an issuer's confidence intervals around either metric do not include the national value for that metric, the issuer would be considered an outlier for that metric within that HCC group.

By the nature of the above formulas, confidence intervals would be wider for issuers with low HCC counts in an HCC group and narrower for issuers with large HCC counts in an HCC group, resolving the inability of the national confidence intervals under the current methodology to be scaled according to each issuer's HCC count in each HCC failure rate group. Simulation results using 2017 benefit year HHS-RADV data seem to suggest that this is the case, with the practical confidence levels for the non-validation rate and the new-found rate becoming nearly indistinguishable from the 95 percent theoretical value and displaying no major trends with regard to sample HCC count (Figure 3.6).

Figure 3.6. Resampling Simulation Comparing Theoretical and Practical Confidence Levels of Current and Binomial Distribution HHS-RADV Methodology by the Number of HCCs in a Sample



As noted above, because we would no longer be using failure rates under this methodology, a new formula for calculating adjustments to risk scores and risk adjustment transfers based on HHS-RADV results would be necessary. Below, Table 3.7 illustrates an example of an adjustment calculation method.

Table 3.7 Potential HHS-RADV HCC Count Adjustment Formulas under the Binomial Distribution Methodology

Non-Validation Group Adjustment	New Found HCCs Group Adjustment
$adj_{NVR} = NVR - NVR_{national}$	$adj_{FndR} = \frac{newFoundHCC_{IVA,national}}{freq_{EDGE,national}} - \frac{newFoundHCC_{IVA}}{freq_{EDGE}}$

As with the group adjustment in the current error estimation methodology, these values may be aggregated first at the enrollee level, then at the issuer level to arrive at a total error rate for the issuer and thereby inform the adjustments to risk scores and risk adjustment transfers.[73] While additional testing will be needed, we believe that the error rates represented by this method will be similar conceptually and in magnitude to the error rates calculated under the current error estimation methodology. However, adoption of this approach would represent a significant change to outlier detection, and would be applicable to all issuers who participate in HHS-RADV for a given benefit year. We are therefore sensitive to the disruptive nature of this option for all issuers of risk adjustment covered plans as we search for alternative options for issuers with low HCC counts. Although we continue to analyze this option, we believe that this option may be the best long-term approach to improve the precision of the outlier detection

[73] See Section 11.3.3 of the 2018 HHS-RADV Protocols at: https://www.regtap.info/reg_librarye.php?i=2904

process and address the inability of the current methodology to scale confidence intervals to provide appropriate results across all HCC counts.

3.2.2.2 McNemar's Test Methodology

The context of the HHS-RADV IVA and SVA processes also prompted HHS to consider a second option based on the binomial distribution: McNemar's test. This test originated as a chi-square test that could be used to test whether bias is present in the disagreement between two measurements of the same dichotomous variable.[74] The below graphic (Table 3.8) may help illustrate this principle as applied to HHS-RADV.

Table 3.8. Simplified Cross-Tabulation of Possible Coding Scenarios for HCCs in EDGE and Audit Data

		Audit Data	
		Absent	Present
EDGE	Absent	$absentHCC$	$newFoundHCC_{IVA}$
	Present	$missingHCC_{IVA}$	$validatedHCC$

This test ignores cases where EDGE and audit data match ($absentHCC$ and $validatedHCC$). When there is a mismatch in coding ($missingHCC_{IVA}$ or $newFoundHCC_{IVA}$), McNemar's test determines whether there is evidence that, when a mismatch between EDGE and audit data has been identified, the mismatch is more likely to fall under $missingHCC_{IVA}$ or under $newFoundHCC_{IVA}$ (that is, EDGE says "present" while audit data says "absent," or EDGE says "absent" while audit data says "present").

The basic concept behind the test may be expressed as asking whether there is evidence that the equation $\frac{missingHCC_{IVA}}{newFoundHCC_{IVA}+missingHCC_{IVA}} = .5 = 50\ percent$ is *not* true. McNemar's test allows for a wide range of coding errors (mismatches) as long as these errors are unbiased.

As with the Binomial Distribution methodology option described in Section 3.2.2.1, under McNemar's test, we would no longer assess issuers based on their failure rates. Instead, we would consider mismatches between EDGE and audit data, ignoring situations in which audit and EDGE data are consistent with one another. We would then calculate a value that represents the rate at which an HCC appears in EDGE, but not in the audit data, given that we know a mismatch has occurred (the non-validated/mismatch ratio).

For any individual HCC, the proportion of mismatches between audit data and EDGE that would represent non-validated HCCs could be expressed as:

[74] Levin, Joel R., and Ronald C. Serlin. "Changing students' perspectives of McNemar's test of change." Journal of Statistics Education 8, no. 2 (2000): 532-541.

$$NV_{mm} = \frac{missingHCC_{IVA}}{newFoundHCC_{IVA} + missingHCC_{IVA}}$$

Whereas the proportion of mismatches between audit data and EDGE that would represent newly found HCCs could be expressed as:

$$Fnd_{mm} = \frac{newFoundHCC_{IVA}}{newFoundHCC_{IVA} + missingHCC_{IVA}} = 1 - NV_{mm}$$

Because these two values are related to one another as described in these formulas, testing NV_{mm} is the same as testing Fnd_{mm}.

To provide an illustrative example of this methodology, an issuer may have the following values for the low failure rate HCC group (Table 3.9). The current measure of failure rate (*GFR*), NV_{mm}, and Fnd_{mm} may all be calculated from this table.

Table 3.9. Example Cross-Tabulation of the Low Failure Rate HCC Group at One Issuer

		Audit Data		
		Absent	Present	Total
EDGE	Absent	6468	20	6488
	Present	25	87	112
	Total	6493	107	6600

$$GFR_i^G = 1 - \frac{Freq_{IVA_i}{}^G}{Freq_{EDGE_i}{}^G} = 1 - \frac{107}{112} = 0.045$$

$$NV_{mm} = \frac{missingHCC_{IVA}}{newFoundHCC_{IVA} + missingHCC_{IVA}} = \frac{25}{20 + 25} = 0.556$$

$$Fnd_{mm} = \frac{newFoundHCC_{IVA}}{newFoundHCC_{IVA} + missingHCC_{IVA}} = \frac{20}{20 + 25} = .444$$

In the current methodology, the *GFR* value would be compared against the lower and upper bounds of the national confidence intervals for the low failure rate HCC group: -0.143 to 0.238. The fact that this value (0.045) is in between these two values signifies that HHS would not consider the value for this issuer to be different from the national mean of the low failure rate HCC group (0.048). Therefore, the issuer in this example would not be considered an outlier. However, under the McNemar's Test methodology, the confidence interval would be approached differently.

As with the *NVR* and *FndR* under the Binomial Distribution methodology, the NV_{mm} (and Fnd_{mm}) would be theoretically distributed according to the binomial distribution and could be assessed for outlier status based on the confidence intervals around issuers' estimates, as determined based on this distribution. However, to obtain a two-sided confidence interval, we would calculate the limits for the McNemar's Test confidence interval as:

$$LL = \frac{2dp + z^2 - \left(z\sqrt{z^2 - \frac{1}{d} + 4dp(1-p) + (4p-2)} + 1\right)}{2(d + z^2)}$$

And the upper limit as:

$$UL = \frac{2dp + z^2 + \left(z\sqrt{z^2 - \frac{1}{d} + 4dp(1-p) - (4p-2)} + 1\right)}{2(d + z^2)}$$

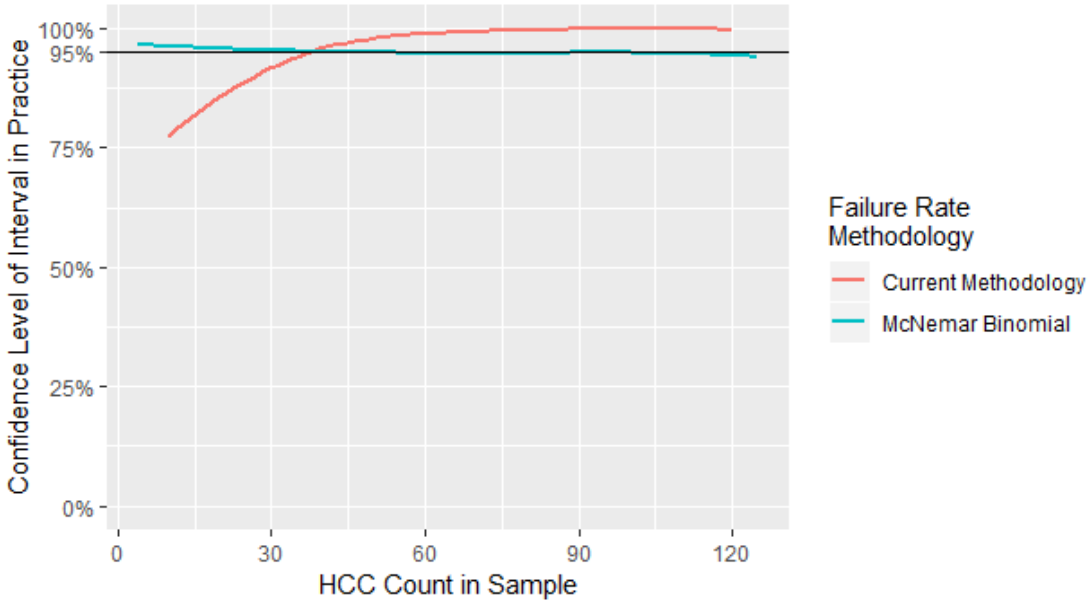
Where:

- d is the denominator of NV_{mm};
- p is the NV_{mm};
- z is the z-value cutoff for a 95 percent two-sided confidence interval: 1.96.

If the issuers' confidence intervals around the NV_{mm} do not include 0.50, the issuer's NV_{mm} rate would be considered an outlier for that HCC group. In the case of the above example, the confidence interval bounds around the NV_{mm}value (0.556) spanned from 0.412 to 0.691. As this value contains the expected value of 0.50, the issuer in this example would not be considered to be an outlier under the McNemar's Test methodology.

The inclusion of the *d* and *p* terms in the above formulas will lead to wider confidence intervals for issuers with lower HCC counts. This reduces the chance that a low-HCC count, typical issuer will be flagged as an outlier. As seen in Figure 3.10, our simulation results using modified national averages based on the 2017 benefit year HHS-RADV data indicate that this method could support scaling confidence intervals to provide appropriate results across all HCC counts.

Figure 3.10. Monte Carlo Simulation Comparing Theoretical and Practical Confidence Level of Current and McNemar HHS-RADV Methodology by the Number of HCCs in a Sample



Like the Binomial Distribution methodology option described in Section 3.2.2.1, the McNemar's test approach would represent a drastic change relative to the current methodology that would apply to all issuers. However, unlike the Binomial Distribution methodology option, which retains conceptual similarity to the current error estimation methodology, the McNemar's Test approach represents a different baseline for comparison and for HHS-RADV adjustments to risk scores and risk adjustment transfers. Issuers' outlier status would no longer be based on their performance relative to one another, but would be based upon the degree by which the frequency of the occurrence of found HCCs in their HHS-RADV sample differs from the frequency of non-validated HCCs in their HHS-RADV sample (i.e. the degree by which these frequencies are not equal). An issuer that fails McNemar's test would have their risk score adjusted to a risk score that represented equal frequencies of found and non-validated HCCs for that issuer, rather than to a risk score that reflected a ratio of found and non-validated HCCs that corresponds to the national mean ratio. The group adjustment for an issuer who fails the test for outlier status could be calculated as:

$$adj_{McNemar} = \frac{missingHCC_{IVA}}{freq_{EDGE}} - \frac{0.5 * (newFoundHCC_{IVA} + missingHCC_{IVA})}{freq_{EDGE}}$$

HHS believes that it is important to note that because issuers would no longer be judged against their peers under this option, and would instead be judged based on a common, universal criterion, it is likely that many more issuers would be flagged as outliers under the McNemar's Test methodology than under the current methodology, which would lead to greater total absolute adjustments to transfers for all issuers (including non-outlier issuers) based on HHS-RADV results.

3.2.3 Alternative Methodologies Using Advanced Techniques Considered

HHS also considered two advanced methodologies that we believed may offer potential improvements on the current error estimation methodology. Although neither of these options appears workable at the present time, we believe that it is beneficial to share them here as they are approaches we considered as part of the development of this white paper.

3.2.3.1 Bayesian Method

In general, Bayesian methods are treated as a separate school of statistical methodology from the classical methods described above. One major advantage of Bayesian methods over classical methods is their ability to update earlier estimates with new data as it comes in. In the HHS-RADV process, these data could be applied in such a way as to gain a narrower interval estimate of an issuer's failure rate with each subsequent year of HHS-RADV. In essence, if an issuer is atypical in their failure rate, and does not take action to improve it, we can narrow-in on an estimate of exactly how atypical that issuer is with each subsequent year of HHS-RADV. New issuers can be incorporated into this process, but the estimates of their failure rates in their initial years participating in the program will be less precise than for issuers who have been participating in HHS-RADV for longer.

Using the precision of their individual failure rate estimates, issuers would be classified as outliers based on our certainty regarding whether their failure rate estimates can be determined to be different from the national mean for a given HHS-RADV year. As a general matter, issuers with smaller HCC counts tend to have less precise estimates than issuers with larger HCC counts. This method would allow us to increase the precision of the low HCC count issuers over time, allowing HHS to better justify adjustments to risk scores and transfer amounts for these issuers with each subsequent year of HHS-RADV data.

We note two major areas in which HHS sees this methodology as impractical. First, Bayesian statistics are less widely understood than the classical statistical methods we are exploring. As such, the HHS-RADV adjustments resulting from a Bayesian method may be less transparent to issuers and other stakeholders. Second, because of the use of prior years' data in the estimate of the current year's failure rate, it may be difficult for atypical issuers to avoid HHS-RADV adjustments to risk scores in the short-term by improving their EDGE coding processes. In other words, if an issuer with a history of poor performance improves, it may take multiple years of HHS-RADV participation after their performance improves for the methodology to reflect that improvement with certainty, possibly leading to the issuer continuing to be flagged as an outlier for several years after that issuer reached a failure rate more typical of its peers.

3.2.3.2 Machine-Learning Method

Based on previous stakeholder comments, we also explored machine-learning as a potential alternative approach for HHS-RADV error estimation. This methodology would largely avoid the limitations of the current methodology by allowing a computer algorithm to decide which issuers were typical or atypical without dictating a particular grouping of HCCs.

In this process, we would apply two unsupervised machine learning algorithms—Isolation Forest and Local Outlier Factor—which we determined to be applicable to the great

variety of HCCs underlying issuers' HCC failure rates. As a result, we would have identified a set of issuers within a single HHS-RADV benefit year dataset that appear as atypical, without the need for any *a priori* constraints imposed by HHS. As such, this technique would allow us to identify issuers whose pattern of failure rates is different from the rest, rather than identify issuers whose overall failure rates are different from the rest.

We could then use a dimension reduction technique to visualize the results of the machine learning algorithms, displaying the variation in issuers' HCC failure rates along two- or three-dimensions. We tested multiple dimension reduction techniques during our exploration, including:

1. Multi-dimensional scaling;
2. T-distributed stochastic neighborhood embedding;
3. Locally linear embedding; and
4. Principal components analysis.

We found that the first and second of these methods were the best for HHS-RADV data, providing a clear delineation between issuers whom the machine learning algorithms identified as typical, and those they identified as atypical.

This machine-learning process has the benefit that we would not be required to define a specific metric and to know how that metric is distributed. Furthermore, it would allow precise control over the proportion of issuers to flag as outliers, as opposed to the current methodology, in which more or fewer than 5 percent of issuers may be flagged per an HCC grouping due to the extremity of their failure rate values or random variation.

However, this option poses several obstacles to a full implementation. First, the method does not intrinsically imply a process by which HHS-RADV results could inform adjustments to risk scores and risk adjustment transfers. Furthermore, we would have difficulty providing a clear explanation as to why a particular issuer was flagged as an outlier. The dimension reduction technique may compute different dimensions year-to-year, and these dimensions may not be readily interpretable by humans, making it very difficult for stakeholders to understand the HHS-RADV results and, more importantly, for issuers to plan or price for expected outcomes of HHS-RADV. Finally, although the algorithms we have explored so far could offer precise control over the proportion of issuers flagged, if we were to utilize this level of control, we would likely need to require that the same proportion of issuers always be flagged as outliers to ensure regulatory consistency and the predictability of issuers' HHS-RADV outcomes year-to-year, even if major improvements were seen in EDGE data quality in subsequent HHS-RADV years.[75] As such, at this stage of our analysis, we do not consider this method viable as a replacement to the current error estimation methodology.

[75] We note, however, that there may be other machine learning algorithms that would allow for the proportion of issuers flagged to vary as EDGE data quality improves nationally. We continue to explore these possibilities.

3.3 Addressing the Influence of HCC Hierarchies on Failure Rate Outlier Determinations

HHS utilizes two sets of medical condition groupings in the HHS-RADV process. The first set—HCCs—originates in the risk adjustment models and is used to aggregate the tens of thousands of standard disease codes used to capture diagnoses into a set of medically meaningful but statistically manageable categories. HCCs in the current HHS risk adjustment models are derived from ICD-9-CM codes that are aggregated into diagnostic groups (DXGs), which are in turn aggregated into broader condition categories (CCs).[76] Then, we apply clinical hierarchies to the CCs, creating subgroupings that contain a set of related or similar medical conditions ranked in order of severity. In the risk adjustment models, if an individual enrollee has more than one CC recorded in EDGE for a given hierarchy, only the most severe of those CCs will be applied for the purposes of risk adjustment. Once hierarchies are imposed, we refer to the codes as HCCs. For example, diabetes diagnosis codes are organized in a Diabetes hierarchy, consisting of three CCs arranged in descending order of clinical severity and cost, from CC 19 *Diabetes with Acute Complications* to CC 20 *Diabetes with Chronic Complications* to CC 21 *Diabetes without Complication*. A person may have diagnosis codes in CC 20 and CC 21, but once hierarchies are imposed, that enrollee would only be assigned the single highest HCC in the hierarchy—HCC 20 *Diabetes with Chronic Complications*. In a typical model recalibration, estimated coefficients of the various HCCs within a hierarchy will ensure that more severe and expensive HCCs within that hierarchy receive a higher risk score than less severe and expensive HCCs. However, in some hierarchies, for various reasons we may constrain coefficients of two or more HCCs to be equal. These reasons may include a "hierarchy violation"—in which the estimated coefficient for an HCC is larger than the coefficient for an HCC above it in the hierarchy—or evidence that it is relatively easy to miscode an HCC as a more severe condition without being detected by medical record review. The Diabetes hierarchy is one such hierarchy where we found it necessary to apply constraints during model recalibration. As such, the three HCCs within the Diabetes hierarchy have been constrained to have the same coefficient in risk adjustment.

Under the current risk adjustment models,[77] there are 127 HCCs, of which 97 HCCs are included among 25 distinct hierarchies. Diagrams and tables of the current hierarchy structure are available in Appendix D.

The other set of medical condition groupings in HHS-RADV is imposed during the error estimation stage of the HHS-RADV process. This set of groupings, the HCC failure rate groupings, is designed to provide a balance between the need to assess the impact of medical coding errors of individual HCCs on risk scores and risk adjustment transfers and the need to assess failure rates on enough HCCs to provide statistically meaningful HHS-RADV results. Furthermore, these groupings are intended to reflect our belief that some HCCs are more difficult

[76] On June 17, 2019, CMS released a paper describing potential HCC updates at: https://www.cms.gov/CCIIO/Resources/Regulations-and-Guidance/Downloads/Potential-Updates-to-HHS-HCCs-HHS-operated-Risk-Adjustment-Program.pdf.
[77] Ibid.

to code accurately than other HCCs, and therefore to provide different national standards based on the level of coding difficulty for a given HCC.

In this HHS-RADV HCC failure rate grouping process, we first calculate the national average failure rate for each HCC individually. HCCs are then ranked in order of their failure rates and split into three groups—a low, medium, and high failure rate group—such that the total occurrence of HCCs in each group nationally is about equal (Table 3.11).

Table 3.11: Number of Unique HCCs in Each HCC Grouping in the 2017 and 2016 BY HHS-RADV Results

	2017 BY: Number of Unique HCCs	2016 BY: Number of Unique HCCs
Low HCC Failure Rate Group	33	33
Medium HCC Failure Rate Group	35	39
High HCC Failure Rate Group	59	55

These HCC failure rate groupings form the basis of the failure rate outlier determination process, with each grouping receiving an individual assessment of outlier status for each issuer. A table of the HCC failure rate groupings for 2017 benefit year HHS-RADV is available in Appendix E.

Based on our experience with the initial years of HHS-RADV, HHS has noticed that in certain situations, these two sets of groupings, the HHS-RADV HCC failure rate groupings and HHS-RA HCC hierarchies, can interact in varying ways that may sometimes lead to misalignments between the HCC failure rate grouping in HHS-RADV and the HCC's hierarchy placement in risk adjustment. The following are examples (which we refer to as "HCC-swapping") of how the hierarchies can interact with HCC failure rate groupings in HHS-RADV:

1. HCCs in the same HHS-RA HCC hierarchy with different coefficients are sorted into different HHS-RADV HCC failure rate groupings: If one HCC is commonly miscoded as another HCC in the same hierarchy in risk adjustment, but the two HCCs are sorted into different HCC failure rate groupings in HHS-RADV, an issuer may be flagged as an outlier in either of the HCC failure rate groupings where one HCC is missing or the other HCC is found.

   For example, HCC 8 *Metastatic Cancer* and HCC 11 *Colorectal, Breast (Age < 50), Kidney, and Other Cancers* are in the same hierarchy in risk adjustment, but for the 2017 benefit year of HHS-RADV, HCC 8 was in the medium HCC failure rate grouping and HCC 11 was in the high HCC failure rate grouping. In validating an enrollee with HCC 8 in HHS-RADV, the IVA or SVA Entity may find that an enrollee with HCC 8 reported in EDGE is not validated as having HCC 8, which is at the top of the HCC hierarchy in risk adjustment, but the enrollee may have been found to have HCC 11 in the issuer's HHS-RADV audit data.

   In this case, HCC 8 would be considered missing in the medium HCC failure rate grouping, and HCC 11 would be considered found in the high HCC failure rate grouping. Other HCCs in the HCC failure rate groupings may then influence the failure rate for that issuer, potentially leading to the issuer being determined to be an outlier in the medium or high HCC failure rate grouping. If the issuer were found to be an outlier in one of the two

failure rate groupings, but not the other, the issuer's HCC failure rate would not represent the difference in risk and costs between these two coefficients in the issuer's HHS-RADV results.

2. HCCs in the same HHS-RA HCC hierarchy with different coefficients are sorted into the same HHS-RADV HCC failure rate grouping: If one HCC is commonly miscoded as another HCC in the same hierarchy in risk adjustment, and the two HCCs are sorted into the same HCC failure rate grouping in HHS-RADV, an issuer may not be flagged as an outlier in that HCC grouping. This may happen because the failure to validate an HCC in HHS-RADV and the discovery of a new HCC in that same HCC failure rate grouping have a net impact of zero on the total final value of an issuer's failure rate in HHS-RADV.

   For example, HCC 35 *End-Stage Liver Disease* and HCC 34 *Liver Transplant Status/Complications* are in the same hierarchy in risk adjustment and were both sorted into the medium HCC failure rate grouping in the 2017 benefit year HHS-RADV results. In validating an enrollee with HCC 35 in HHS-RADV, the IVA or SVA Entity may find that an enrollee with HCC 35 reported in EDGE is not validated as having HCC 35, but the enrollee may have been found to have HCC 34 in audit data.

   In this case, not validating HCC 35 and finding HCC 34 in the same HCC grouping in HHS-RADV would, when taken together, have no net impact on the issuer's HCC group failure rate. In essence, for the purposes of the calculation of the failure rate, it appears that there is no difference between HCC 34 and HCC 35, even though these two HCCs have different coefficients in risk adjustment. Because these HCCs have different risk and costs, the inability for the issuer's HCC failure rate to identify that an individual has HCC 34 rather than HCC 35 results in an inability to represent the difference in risk and costs between these two coefficients in the issuer's HHS-RADV results.

3. HCCs in the same HHS-RA HCC hierarchy with constrained coefficients are sorted into different HHS-RADV HCC failure rate groupings: Another way in which HCC failure rate groupings and hierarchies may interact is a compounding of the first example—the sorting of HCCs from the same hierarchy into different failure rate groups—and constrained coefficients in risk adjustment. In HHS-RADV, if two HCCs in the same hierarchy have coefficients that are constrained are sorted into different HCC failure rate groupings, a sufficient miscoding of one HCC for the other may lead to the issuer being identified as a positive outlier in one HCC failure rate grouping or a negative outlier in another HCC grouping, despite there being no difference in risk score due to the coding error.

   For example, HCC 54 *Necrotizing Fasciitis* and HCC 55 *Bone/Joint/Muscle Infections/Necrosis* share a hierarchy in risk adjustment and have their risk score coefficients constrained to be equal, but for 2017 benefit year HHS-RADV, HCC 54 was in the high failure rate HCC grouping, while HCC 55 was in the medium failure rate HCC grouping. In validating an enrollee with HCC 54 in HHS-RADV, the IVA or SVA Entity may find that an enrollee with HCC 54 reported in EDGE is not validated as having HCC 54, but the enrollee may have been found to have HCC 55 in audit data.

In this case, when taken together with the issuer's other HHS-RADV results, these HCCs with the same coefficients could contribute to an issuer's failure rate in either the high failure rate grouping or the medium failure rate grouping, even though the HCCs do not have different risk scores and an adjustment to risk score is not conceptually warranted.

4. HCCs in the same HHS-RA HCC hierarchy with constrained coefficients are sorted into the same HHS-RADV HCC failure rate grouping: HCC groupings and hierarchies may interact in a fourth way. If two HCCs share a hierarchy with constrained coefficients in risk adjustment and are sorted into the same HCC failure rate grouping in HHS-RADV, and an enrollee has the first HCC in the HCC group but this HCC fails to be validated in HHS-RADV while another HCC in that HCC group is newly discovered for that enrollee during HHS-RADV, the missing and found HCCs will have a net impact of zero on both the failure rate and risk score.

   For example, HCC 20 *Diabetes with Chronic Complications* and HCC 19 *Diabetes without Complications* share the same hierarchy and have their coefficients constrained to be equal. In the 2017 benefit year HHS-RADV results, HCC 19 and HCC 20 are both in the low HCC failure rate grouping. If an enrollee is not validated as having HCC 20 during the 2017 benefit year HHS-RADV audit procedures and is instead found to have HCC 19, the issuer's failure rate is unaffected by the change from one to the other HCC and no change in risk score would be applied as a result, nor would a change in risk score be conceptually warranted.

We have performed an initial review of the occurrence of these scenarios in the 2017 benefit year HHS-RADV results. Of all the HCCs in EDGE that were not validated in the audit data, about 1/8th represent HCCs that IVA or SVA auditors coded as different HCCs within the same hierarchy. Of the HCCs that were newly found in the audit data – that is, they were not recorded in the original EDGE data – around 1/3rd represent HCCs that were newly found because they were originally reported on EDGE as a different HCC in the same hierarchy. However, we note that these occurrences are distributed among the four scenarios previously described and, therefore, for many issuers, would be unlikely to impact whether they were an outlier in an HCC failure rate grouping.

The methodologies described in this chapter provide varying levels of improvement in the precision of outlier detection and scaling the confidence intervals used to determine outlier status to better account for variation in HCC counts. However, the influence of the interaction between HCC hierarchies and HCC failure rate groups in error estimation and outlier determination persists throughout all of the techniques previously described in this chapter. As such, we are in the preliminary stages of exploring HHS-RADV methodology alternatives that would help mitigate the cases where the misalignment between the HCC grouping in HHS-RADV and the HCC's hierarchy placement in risk adjustment occurs, with the goal to better account for HCCs that are miscoded within an HCC hierarchy when there is a difference in risk and costs between the HCCs. The options described in the next sections include an approach involving assessment of ordinal-by-ordinal relationships and an approach that bases outlier status on the distribution of enrollee-level risk scores, rather than issuer-level HCC failure rates. Although we describe these

options in this paper, we are continuing to generate and assess different options, and are interested in comments on whether there are other options that we should consider to refine the HHS-RADV methodology to better account for HCCs that are miscoded in the same hierarchy when there is a difference in risk and costs between the HCCs.

3.3.1 Ordinal-by-ordinal relationships as Applied to HHS-RADV

The first option that we explored to better account for HCCs that are miscoded within an HCC hierarchy when there is a difference in risk and costs between the HCCs is an approach involving assessment of ordinal-by-ordinal relationships. This type of assessment provides a single test that investigates all of the ways in which HCCs that share a hierarchy can be miscoded in EDGE: missing from audit data, newly discovered in audit data, and swapped to a different HCC in audit data. This approach would test the correspondences between EDGE and audit data for HCCs that share a hierarchy all at once, avoiding the situation in which an issuer may be flagged as an outlier multiple times for various HCCs within the same hierarchy.

As a class of statistics, ordinal-by-ordinal relationships indicate the degree to which higher values on one ordinal variable correspond to higher values on another ordinal variable.[78] Applying this framework to HHS-RADV would involve redefining how we identify which HCC from a hierarchy of HCCs an enrollee has. Currently, we identify each HCC with a separate yes or no question; that is, "does this enrollee have HCC X from Hierarchy A? Does this enrollee have HCC Y from Hierarchy A?" and so on. To implement a test based on ordinal-by-ordinal relationships, we would represent these separate questions as a single question regarding which HCCs within a hierarchy a participant had. For example, the following three HCCs all share a hierarchy:

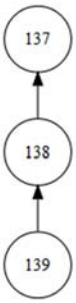
- HCC 137 *Hypoplastic Left Heart Syndrome and Other Severe Congenital Heart* Disorders
- HCC 138 *Major Congenital Heart/Circulatory Disorders*
- HCC 139 *Atrial and Ventricular Septal Defects, Patent Ductus Arteriosus, and Other Congenital Heart/Circulatory Disorders*

In this hierarchy, HCC 137 supersedes HCC 138, which supersedes HCC 139, indicating that HCC 137 is the most severe and expensive of these HCCs[79] (Figure 3.12).

[78] E.g. Normal Cliff and Ventura Charlin, "Variances and Covariances of Kendall's Tau and Their Estimation," *Multivariate Behavioral Research* 26, no. 4 (1994): 693-707. DOI: 10.1207/s15327906mbr2604_6.

[79] In the context of the HHS risk adjustment models, this means that an enrollee who had been coded as having 137 and 138 would only receive the risk score associated with 137, the more severe of the two.

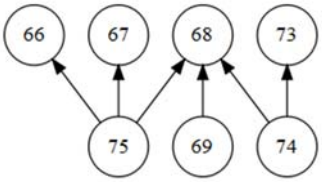
Figure 3.12. Structure of the Heart Abnormality HCC Hierarchy



An ordinal recoding of these variables would be to call them one variable, "congenital heart abnormalities" with four levels in order of severity: 0 = None-in-hierarchy; 1 = HCC 139; 2 = HCC 138; 3 = HCC 137. The EDGE version of this ordinal variable and the audit data version of the ordinal variable could then be compared using a non-parametric correlation coefficient such as Kendall's Tau.

However, this approach presents a major challenge in that some HCC hierarchies have rather complex structures, and we need to investigate further how to recode them as ordinal variables. One such complex hierarchy is seen in Figure 3.13.

Figure 3.13 Structure of the Blood and Immune Disorder Hierarchy



According to this hierarchy, 75, 69 and 74 appear to be on the same level, as do 66, 67, 68, and 73, but several of these HCCs have no defined supersession relationship; for example, 75 and 73. It is therefore difficult to put these values in any particular order. For this reason, we have concerns about this option being a viable alternative to the current methodology.

3.3.2 Assessing Outlier Status based on Risk Score Directly

To develop an option that does not require hierarchies to have any particular structure, we took into consideration that MA-RADV uses a methodology based on assessing the statistical significance of errors in Medicare Advantage (MA) payments directly, rather than through a measure of the frequency of HCC validation failures. We initially expected to use a similar methodology that could create a confidence interval around an issuer's total PLRS. However, we determined that several factors weighed against a close replication of the MA-RADV methodology for HHS-RADV. Two of these factors include:

- Supporting Market Predictability: Our desire to reduce the number of issuers facing adjustments to promote market stability in budget neutral risk pools by adjusting issuers' risk scores based on significant deviation from a national average non-validation value; and

- Different Diagnosis Patterns: A lower frequency of diagnoses in risk adjustment covered plans than MA for common conditions[80] resulting in a greater impact of errors when, for valid clinical reasons, an HCC is difficult to code.

These factors are currently addressed in HHS-RADV through the creation of national confidence intervals around the weighted mean failure rate and the establishment of the three HCC failure rate groups. If HHS-RADV were to transition to a new methodology based on measuring errors in risk scores directly, it would need to continue to address these factors.

Due to input from stakeholders that the current methodology does not take into consideration the impact of HCC hierarchies on outlier detection, we are considering an approach that could create confidence intervals around estimates related to each issuer's per-enrollee average PLRS, with the width of these confidence intervals being defined by the theoretical sampling distribution of that issuer's enrollee risk scores for the applicable benefit year. We have developed a draft approach that may satisfy the above factors, as well as issuer concerns about the impact of HCC hierarchies on outlier detection. This approach could include the following key components:

- Assign HCCs to groups (analogous to HCC failure rate groups in the current methodology) by whole hierarchies, rather than by individual HCCs. This would eliminate all instances of HCCs that are in the same HHS-RA HCC hierarchy being sorted into different HHS-RADV HCC failure rate groupings (HCC-swapping examples 1 and 3 above) regardless of whether the coefficients within that hierarchy have been constrained, and would ensure that swaps within a hierarchy are not counted separately in different groupings.[81]

- Create these groups according to the difference in total risk score between EDGE and audit data for the HCCs in each hierarchy relative to the EDGE risk score, rather than by difference in HCC count relative to the EDGE HCC count (the current approach). That is, create "hierarchy risk score discrepancy" groups instead of HCC failure rate groups. This would eliminate the effects of HCC-swapping examples 2 and 4 by ensuring that swaps are always credited or debited in proportion to their effect on risk score when an issuer is determined to be an outlier in one or more "hierarchy risk score discrepancy" groups. Examples of formulas for determining the relative difference by which to rank the hierarchies could be as follows:

$$EDGE.RS_{hier} = \sum_{\boldsymbol{metal}} \sum_{\boldsymbol{hcc}} \left(EDGE.Freq_{metal,hcc,hier} * coef_{metal,hcc} \right)$$

[80] Approximately 80 percent of enrollees in risk adjustment covered plans had zero HCCs reported through EDGE servers (https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/Summary-Report-Risk-Adjustment-2018.pdf), whereas 49 percent of MA enrollees had zero HCCs reported (https://www.ahip.org/wp-content/uploads/Wakely-2020-Medicare-Advantage-Adv-Notice-and-Risk-Model-Impact-Report-2.28.2019-1.pdf).

[81] Alternatively, we could consider eliminating the use of groupings based on risk score discrepancy/failure rate under this option.

$$Audit.RS_{hier} = \sum_{\boldsymbol{metal}} \sum_{\boldsymbol{hcc}} \left(Audit.Freq_{metal,hcc,hier} * coef_{metal,hcc} \right)$$

$$relDiff_{hier} = \frac{EDGE.RS_{hier} - Audit.RS_{hier}}{EDGE.RS_{hier}} = 1 - \frac{Audit.RS_{hier}}{EDGE.RS_{hier}}$$

Where:

- *EDGE.RS$_{hier}$* is the total risk score in EDGE for all HCCs in a hierarchy;
- *EDGE.Freq$_{metal,hcc,hier}$* is the frequency in EDGE for each HCCs in a hierarchy at each metal level, where *metal, hcc, hier* are the indexes for insurance metal level, HCCs and hierarchies;
- *coef$_{metal,hcc,}$* is the coefficient for a given HCC from RA for each metal level;
- *Audit.RS$_{hier}$* is the total risk score in audit data for all HCCs in a hierarchy;
- *Audit.Freq$_{metal,hcc,hier}$* is the frequency in audit data for each HCCs in a hierarchy at each metal level;
- *relDiff$_{hier}$* is the relative difference between the audit and EDGE risk scores for a given hierarchy, where *metal, hcc, hier* are the indexes for insurance metal level, HCCs and hierarchies.

- Determine outlier status by comparing the following values (list items *a* and *b,* below):
  a. the average difference between the enrollee-level risk score in EDGE and the enrollee-level risk score in the audit data for each "hierarchy risk score discrepancy" group for each issuer. An example of formulas that could define these values are as follows:

$$EDGE.RS_{e,i,G} = \sum_{\boldsymbol{hcc}} \left(EDGE.Coded_{e,i,hcc,G} * coef_{metal,hcc} \right)$$

$$Audit.RS_{e,i,G} = \sum_{\boldsymbol{hcc}} \left(Audit.Coded_{e,i,hcc,G} * coef_{metal,hcc} \right)$$

$$diff_{e,i,G} = EDGE.RS_{e,i,G} - Audit.RS_{e,i,G}$$

$$meanDiff_{i,G} = \frac{\sum_e diff_{e,i,G}}{n_i}$$

Where:

- *EDGE.RS$_{e,i,G}$* is the total risk score in EDGE for an enrollee *e* for issuer *i* for group *G;*
- *EDGE.Coded$_{e,i,hcc,G}$* has values of 1 or 0, representing the presence or absence of a given HCC in EDGE for an enrollee *e* for issuer *i* for group *G;*
- *Audit.RS$_{e,i,G}$* is the total risk score in audit data for an enrollee *e* for issuer *i* for group *G;*
- *Audit.Coded$_{e,i,hcc,G}$* has values of 1 or 0, representing the presence or absence of a given HCC in audit data for an enrollee *e* for issuer *i* for group *G;*

- $diff_{e,i,G}$ is the difference between the audit and EDGE risk scores for an enrollee *e* for issuer *i* for group *G*;
- $meanDiff_{i,G}$ is the average enrollee-level difference for each issuer *i* for group *G*;
- n_i is the total number of enrollees in the RADV sample for an issuer *i.*

b. The national average difference between enrollee-level risk score in EDGE and audit data for that "hierarchy risk score discrepancy" group, which could be defined as:

$$natMeanDiff_G = \frac{\sum_i \sum_e diff_{e,i,G}}{N}$$

Where:

- $natMeanDiff_G$ is the national average enrollee-level difference across all issuers for group *G*;
- *N* is the total number of enrollees sampled across all issuers nationally for HHS-RADV.

- This type of approach would not use national confidence intervals based on issuer-level failure rates. Instead, this approach would calculate the standard error as the national standard deviation of the difference between the enrollee-level risk scores in EDGE and in audit data divided by the square root of each issuer's sample size. Then, we would use this standard error in the construction of issuer-specific confidence intervals. Formulas for this step could be as follows:

$$natDiffSD_G = \sqrt{\frac{\sum_i \sum_e \left(diff_{e,i} - natMeanDiff_G\right)^2}{N-1}}$$

$$diffSE_{i,G} = \frac{natDiffSD_G}{\sqrt{n_i}}$$

$$95\%\ CI_{i,G}: meanDiff_{i,G} \pm 1.96 * diffSE_{i,G}$$

Where:

- $natDiffSD_G$ is the standard deviation of all enrollees' difference values across all issuers nationally for a group *G;*
- $diffSE_{i,G}$ is the standard error of the average difference at an issuer *i*;
- $95\%\ CI_{i,G}$ is the confidence interval for the average difference in risk score for an issuer *i,* centered on that issuer's average difference.

- Under this approach, an issuer's outlier status would then be determined according to whether the issuer's confidence interval ($95\%\ CI_{i,G}$) captured the national mean difference $natMeanDiff_G$ for a given group *G*. If the national mean is outside the bounds of the issuer's confidence interval, that issuer would be considered an outlier for that "hierarchy risk score discrepancy" group.

Although we have preliminary evidence that this methodology may be a viable option to address HCC-swapping scenarios while satisfying the factors for HHS-RADV described above,

further analysis is needed to confirm the ability of this methodology to identify true outliers and to further assess the impact this methodology would have on HHS-RADV adjustments to transfers.

3.4 Summary of Approaches Detailed in this Chapter

At the present time, based on our current analysis of available data, the information that HHS has compiled thus far indicates that, of the basic modifications to the current methodology described in Section 3.2.1 and the alternative methodologies described in Section 3.2.2, the Binomial Distribution methodology (Section 3.2.2.1) is the most viable long-term solution to refine our outlier detection methodology to more precisely identify true outliers. However, we intend to delve deeper into the risk-score-based methodology (Section 3.3.2) as a way of addressing the impact of HCC hierarchies on outlier detection, and as we continue to explore other options, including consideration of stakeholder comments on this paper, our assessment of the most viable long-term approach may change. In the next chapter, we also consider an alternative option to address feedback from stakeholders about the impact of found HCCs in the calculation of error rates. We are interested in comments on the various options described in this chapter and other methods we should consider to address these issues.

4. Error Rate Calculation

This chapter focuses on the calculation of an outlier issuer's error rate as described in Section 1.2.3 of this paper and whether adjustments to this calculation are needed in cases where the outlier issuer is only slightly outside of the confidence interval for one or more HCC groups, as well as cases where a negative error rate outlier issuer also has a negative failure rate. The first section of the chapter reviews the key factors used in an issuer's error rate calculation. The second section discusses the differences in the types of outliers. The third section discusses the observation of issuers likely to be outliers based on the 2017 benefit year HHS-RADV results. The fourth section reviews how the current error rate adjustment calculation for outliers creates a "payment cliff", then analyzes alternative options to calculate an issuer's error rate to mitigate the "payment cliff" effect. The fifth section discusses negative error rate outliers with negative failure rates with an option to constrain those error rate calculations, and the sixth section considers other options to adjust the error rate calculation beyond adjusting for the "payment cliff." In response to stakeholder feedback, HHS has a particular interest in examining ways to mitigate the impact of the "payment cliff," as well as cases where an outlier issuer has a negative failure rate.

4.1 Key Factors Used in the Error Rate Calculation

As described in Section 1.2.3 of this paper, the calculation of an error rate for an outlier issuer depends on a number of factors. These factors include the frequency of HCCs in the issuer's enrollee sample for the HCC group that was validated by the IVA or SVA, as applicable, the frequency of HCCs in the issuer's enrollee sample for the HCC group in EDGE, the issuer's HCC group failure rates, the national metrics determined for HHS-RADV for that benefit year, and the issuer's sampled enrollee-level original and adjusted risk score.

To calculate the issuer's error rate, the issuer's HCC group failure rates are first calculated by the rate of HCCs validated by the IVA or SVA, as applicable, versus the rate of HCCs on EDGE for the issuer's enrollee sample subtracted from 1. Then, if an issuer's failure rate for an HCC group is determined to be an outlier, that HCC group failure rate is used to determine the issuer's group adjustment factor. The issuer's group adjustment factor for an HCC group is calculated based on the issuer's failure rate and the distance of that failure rate from the weighted mean HCC group failure rate. Once the issuer's group adjustment factor has been calculated for that HCC group, that group adjustment factor is applied directly to sampled enrollee-level HCC risk score factors to calculate the issuer's error rate. Then, that error rate is applied to the issuer's PLRS and results in adjustments to risk adjustment transfers for the applicable state market risk pool.

4.2 Differences in Types of Outliers

HHS-RADV uses a two-sided confidence interval to determine outliers. This approach means that there are both upper and lower bound outliers for each HCC group. An upper bound outlier is a "positive error rate outlier" whereby the issuer's failure to validate the HCCs in its HHS-RADV sample was worse than the confidence interval around the national failure rate for one or more HCC groupings. A lower bound outlier, on the other hand, is a "negative error rate outlier"

whereby the issuer's failure to validate the HCCs in its HHS-RADV sample was better than the national confidence interval for one or more HCC groups. If the error rate is positive, the issuer's PLRS are adjusted downward by the adjustment rate, which results in a higher risk adjustment charge, lower risk adjustment payment, or a shift in the issuer's transfer amount from a payment to a charge, assuming no adjustments to other issuers' PLRS in the same state market risk pool. If the error rate is negative, the issuer's risk scores are adjusted upward by the adjustment rate, which results in a lower risk adjustment charge, higher risk adjustment payment, or a shift in the issuer's transfer amount from a charge to a payment, assuming no adjustments to other issuers' risk scores in the same state market risk pool. Issuers that are outliers in more than one HCC failure rate group have one error rate that is calculated based on all of the HCC groups in which the issuers are an outlier.

Within the group of negative error rate outliers, there is a subgroup of issuers that are negative error rate outliers with negative failure rates. As described in the previous chapter, negative failure rates can occur in HHS-RADV when the audit data contains more HCCs in an HCC group than were recorded on EDGE. Between the 2016 benefit year and 2017 benefit year HHS-RADV results, there was an increase in the number of issuers that were negative error rate issuers with negative failure rates for all HCC groups. We discuss this subgroup of outliers later in this chapter.

Because the current methodology identifies and adjusts for outliers based on a 95 percent confidence interval for each of the three HCC groupings, in any given benefit year, the majority of issuers that participate in HHS-RADV will likely not be outliers and will receive an error rate of zero and no adjustment to their risk score(s). These non-outlier issuers' results are within the confidence intervals of the national HCC group failure rates, but their risk adjustment transfers could nevertheless be adjusted due to other outlier issuers in their state market risk pool(s) due to the budget-neutral nature of the HHS-operated risk adjustment program.

4.2.1 Outlier Observations

In reviewing the 2017 benefit year HHS-RADV results, we observed certain patterns about the distribution of states, issuers, and market risk pools and the prevalence of outlier status in the HHS-RADV results. First, we looked at the difference in outliers between markets. Because there are generally more issuers participating in the small group market risk pools than the other state market risk pools for the 2017 benefit year of HHS-RADV, more small group market issuers were identified as outliers than individual, catastrophic or merged market issuers and saw their risk scores adjusted as a result of HHS-RADV. This resulted in a higher number and level of risk adjustment transfer changes in the small group market risk pools compared to the individual, catastrophic, and merged market risk pools. For the 2017 benefit year, the individual and catastrophic risk pools experienced more issuers leaving the market risk pools than the small group market risk pools. Although issuers that exited the individual and catastrophic risk pools were more likely to be positive error rate outliers, they also tended to have low market shares as they exited the markets and therefore, their error rates when applied to risk scores tended to have a low impact on risk adjustment transfers as shown in Appendix B. Starting with the 2018 benefit year of HHS-RADV, we will adjust for exiting issuers only if the exiting issuers are

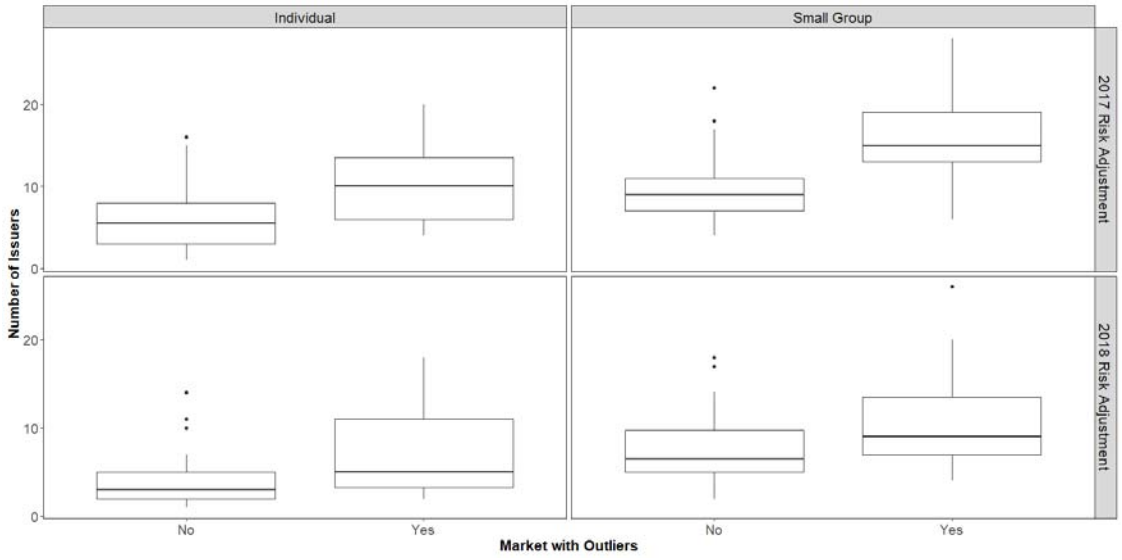
positive error rate outlier issuers, which should limit the number of adjustments being made for exiting issuers.[82]

Second, as we previously stated in the 2016 benefit year HHS-RADV Results memo, based on the empirical failure rate distribution of all issuers in the 2016 benefit year HHS-RADV data, we expected that outliers resulting in positive error rates would be more prevalent than outliers resulting in negative error rates. We found this expectation was supported by 2017 benefit year HHS-RADV results. We also found that in both the 2016 and 2017 benefit years, negative error rate outliers tended to have smaller error rates than positive error rate outliers. We also expect that as issuers gain experience with HHS-RADV, issuers' failure rates will improve, which would result in narrower confidence intervals. These narrower confidence intervals signify a more limited range of failure rates among issuers, resulting in less distance to the weighted means used to calculate the group adjustment factor and by extension lower error rates. We saw this trend between the 2016 and 2017 benefit year HHS-RADV national metric results and anticipate it will continue.

Third, issuers that are outliers in multiple HCC groups were typically outliers in the same direction for each HCC group in benefit year 2017 HHS-RADV results. This trend was also generally observed in the 2016 benefit year HHS-RADV results.

Fourth, in general, the state market risk pools with at least one outlier have a larger number of issuers, compared to states without any outliers. Specifically, there was no relationship between the number of issuers and the percent of issuers being outliers. Likewise, a state with a smaller numbers of issuers participating in a market risk pool was less likely to have an outlier.

Figure 4.1: Comparing the Number of Issuers in a Market to Outliers Being in the Market



[82] 84 FR at 17503-17504.

Lastly, we found in the 2017 benefit year HHS-RADV results that smaller issuers were more likely to be identified as outliers than large issuers. However, this could follow from a major distinction between the 2016 and 2017 benefit year HHS-RADV dataset, in that issuers with $15 million or less in premiums did not have to participate in the 2016 benefit year of HHS-RADV (the second pilot year), but were required to participate in 2017 benefit year HHS-RADV (the first non-pilot year). As these issuers did not have the same experience with the HHS-RADV program in past benefit years as larger issuers, they may not have adapted their medical coding practices, provider engagement, and other factors to the same degree or in the same way as issuers who participated in the HHS-RADV pilot for the 2016 benefit year. The continued, periodic participation in the HHS-RADV program requirement captured in the materiality threshold at 45 C.F.R § 153.630(g)(2) that applies beginning with the 2018 benefit year HHS-RADV will allow us to further evaluate whether smaller issuers will be more likely to be identified as outliers in the long term.

4.3 Application of Thresholds Under the Current methodology

When using a methodology built upon the determination of outliers and a rate of adjustment for those outliers, thresholds are used. In the case of the current methodology, those thresholds are used to determine whether the issuer is an outlier and to determine the error rate that will be used to adjust risk scores and transfers as a result of those outlier issuers' HHS-RADV results. As previously discussed, 1.96 standard deviations on both sides of the confidence interval from the weighted HCC group means are the thresholds currently used to determine whether the issuer is an outlier and the weighted HCC group mean is the threshold used to determine the rate of adjustment. In practice, these thresholds mean that an issuer with failure rates outside this 1.96 range is deemed an outlier and sees an adjustment to its risk score, while an issuer with failure rates inside this 1.96 range sees no adjustment to its risk score. This policy means that all outlier issuers are treated the same in the calculation of their error rates regardless of their relative distance from the confidence interval.

Because the current thresholds used to calculate issuers' error rates are based on the difference between their failure rates and the group weighted mean failure rates, stakeholders have expressed concerns that the current error estimation methodology results in issuers that are just outside of the confidence intervals receiving an adjustment to their risk score, even though they are not significantly different from the issuers just inside the confidence intervals who receive no adjustment to their risk score, creating a "payment cliff" or "a leap frog effect". For example, an issuer with a low HCC group failure rate of 23.9 percent would be considered an upper bound, positive error rate outlier for that HCC group based on the 2017 benefit year national failure rate statistics because the upper bound confidence interval for the low HCC group is 23.8 percent. That issuer's group adjustment factor would be calculated based on the difference between the weighted low HCC group mean of 4.8 percent and the issuer's 23.9 percent failure rate for that HCC group. Under this example, the issuer's group adjustment factor would be 19.1 percent, and that group adjustment factor would be applied to the enrollee-level HCC risk score factors in the issuer's sample population to calculate the error rate. At the same time, another issuer with a similar low HCC group failure rate of 23.7 percent would receive no adjustment to its risk score as a result of HHS-RADV. While this result is due to the nature of

establishing and using a threshold, some stakeholders have argued for limits to the adjustment rate threshold applied to outlier issuers. For example, stakeholders have recommended limits that include calculating error rates based on the position of the confidence interval for the HCC group and not on the position of the weighted mean for the HCC group. Others have recommended not adjusting issuers' risk scores in case of negative error rate issuers to limit the impact of these adjustments on issuers who are not determined to be outliers.

As discussed in prior rulemakings, we have concerns about only adjusting issuers' risk scores for positive error rate outliers. However, we recognize that changing the calculation and application of an outlier issuer's error rate may be appropriate if the outlier issuer is not statistically different from the issuers within the confidence intervals. Thus, our main goal for considering changes to the calculation of the error rate would be to mitigate the "payment cliff" for situations where issuers may be close the confidence intervals and are not substantially different than those issuers inside the confidence intervals. We also discuss in this chapter an option that could change the calculation of the error rate for negative error rate outliers that have negative failure rates.

4.4 Alternative Options to Calculate the Error Rate and their Impact

This section discusses options to revise the current calculation of an outlier issuer's error rate for cases where the outlier issuer is only slightly outside of the confidence intervals for the HCC group. To address the "payment cliff" issue for these issuers, we have considered several options to revise the thresholds used in the error rate calculation to smooth the calculation of the group adjustment factor, including reverting back to the original error estimation methodology, adjusting to the confidence intervals, only adjusting for positive error rate outliers, and several options to apply a sliding scale adjustment factor. While we have looked at some of these options using the 2016 benefit year results, we generally tested these options using the 2017 benefit year HHS-RADV results. One challenge that we ran into in testing these policy options is that an issuer's transfer impact usually occurs for more than one reason. Because of the complexity of this issue, we intend to continue to test our results as future benefit years of data become available, along with other policy options being considered. The following subsections describe and explain our consideration of these options.

4.4.1 Original Error Estimation Methodology

The original error estimation methodology finalized in the 2015 Payment Notice and discussed in Section 1.2.2 of this paper would have adjusted almost all issuers' risk scores for every error identified as a result of HHS-RADV. The adjustments under this methodology would have used the issuer's corrected average risk score to compute an adjustment factor. The adjustment factor would have been based on the ratio between the corrected average risk score and the original average risk score. After taking into consideration the final IVA (or SVA as appropriate) results, we would have calculated the estimated adjusted total population risk score compared to the EDGE total population risk score, and derived a point estimate of the risk score error rate for each issuer based on the original error estimation methodology. We would have calculated adjustments for all issuers with error rates significantly different than zero using a 95 percent confidence interval. In making these calculations for purposes of this paper, we generally used the 2017 benefit year HHS-RADV results to simulate what the risk adjustment transfer

impacts would have been under the original methodology and compared them to the risk adjustment transfer impacts under the current error estimation methodology in Appendix B.[83]

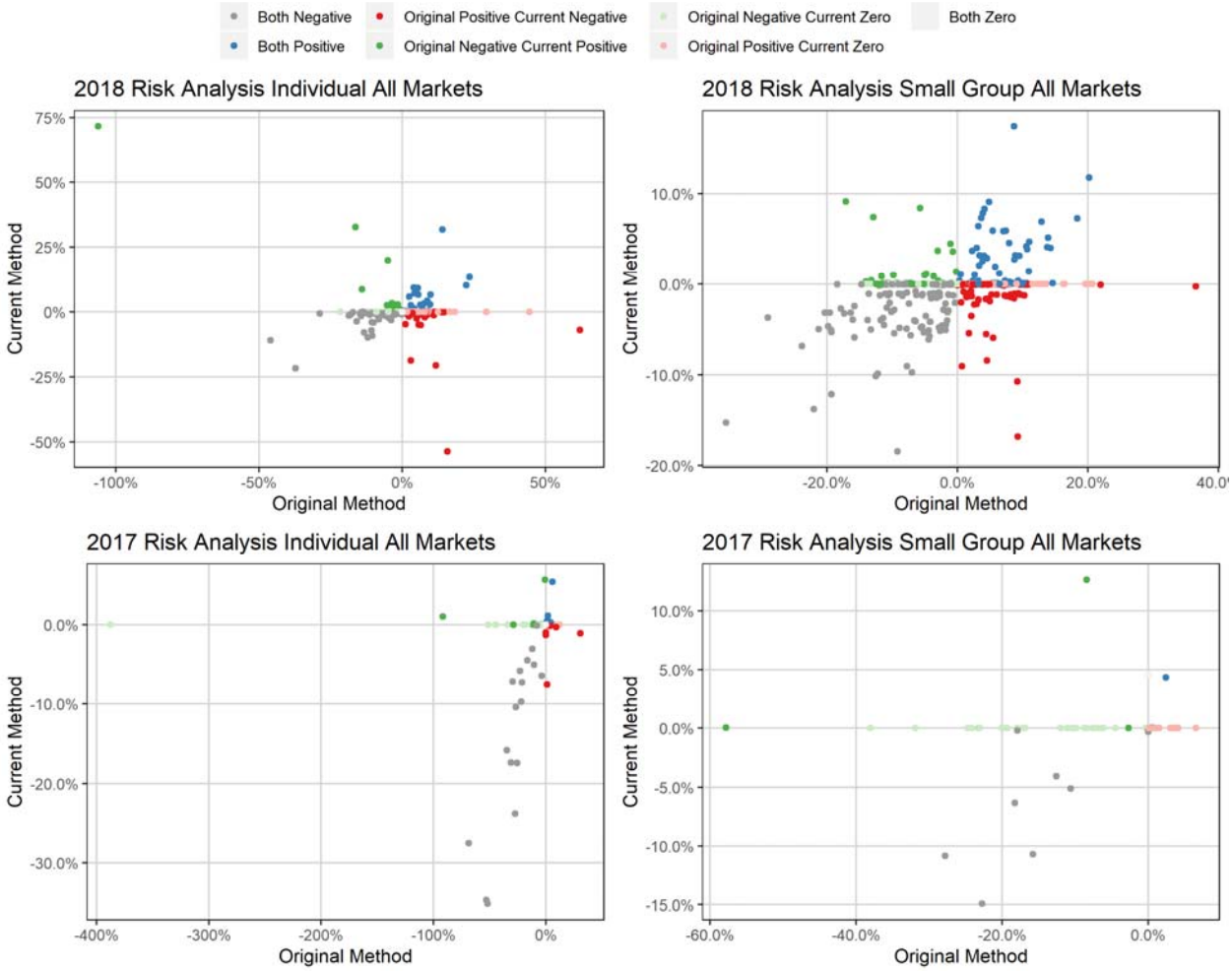
In comparison to the current methodology, this option may be attractive for those issuers who believe that they will typically have HCC group failure rates below the weighted mean that are not low enough to be flagged as a negative error rate outlier for any HCC group under the current methodology. In theory, this option could prevent the "payment cliff," as every issuer's failure to validate an HCC would be taken into account in calculating an issuer's failure rate. Then, assuming that all issuers had some level of failure in the state market risk pool, the error rates being applied in the state market risk pool would be equaled out to some extent because all issuers would see their respective risk scores adjusted, thereby reducing the probability for a "payment cliff".

However, based on our testing of the original error estimation methodology for purposes of this white paper, we did not find this to be the case. Instead, our analysis found that the actual impact of HHS-RADV results on individual issuers' risk adjustment transfers is complex and individual issuers' error rates could decrease or increase under the original methodology in comparison to the current methodology, resulting in larger transfer changes. For example, as shown in Appendix B, we found that applying the 2017 benefit year HHS-RADV results to the original methodology to the small group market risk pools would have resulted in a 121.17 percent change in risk adjustment transfers after HHS-RADV compared to the total risk adjustment transfers before HHS-RADV (in comparison to a 29.81 percent change in transfers under the current methodology).

Additionally, we found that applying the original methodology generally created a more severe "payment cliff," since the majority of adjusted issuers with failure rates significantly different from zero had their original failure rates applied without the benefit of subtracting the weighted mean difference that is used under the current methodology. For these reasons, the risk adjustment transfers move in unexpected ways. Figure 4.2 shows the results of our testing of issuers' 2017 and 2018 risk adjustment transfer change over premiums using the 2017 benefit year HHS-RADV results under the original methodology compared to the current methodology. As shown, almost all issuers would have been adjusted and many issuers would have been negatively impacted under the original methodology compared to the current methodology.

[83] See infra notes 100 and 101 in Appendix B.

Figure 4.2: Comparing Issuers' Transfer Change over Premium for the Original and Current Error Estimation Methodologies on 2017 Benefit Year HHS-RADV Results



Beyond a higher proportion of issuers' risk scores being adjusted under the original methodology, the estimated risk adjustment transfer change using the original methodology was more than four times higher across all markets than the risk adjustment transfer changes under the current methodology as seen in Appendix B.

We also continue to believe that some variation and error should be expected in the compilation of data for risk scores because provider documentation of enrollee's health status varies across provider types and groups. Adjusting almost all issuers for every failure found in the HHS-RADV process, as was the case with the original methodology, does not take into consideration any expected variation and errors. In addition, as detailed in the prior section, we have a strong desire to reduce the number of issuers facing adjustments to promote market stability by adjusting issuers' risk scores based on significant deviation from a national average non-validation value, rather than adjusting for every failure (regardless of the magnitude of the error).

4.4.2 Only Adjusting to Confidence Intervals

As previously discussed, another option suggested by some stakeholders to address the "payment cliff" is to modify the error rate calculation and no longer calculate an outlier issuer's group adjustment factor using the threshold of the distance to the weighted HCC group mean. Instead, under this option, the issuer's group adjustment factor for its error rate calculation would be calculated using the HCC group confidence interval. This option could ensure that outlier issuers with failure rates just outside of the confidence intervals, as well as the outlier issuers with failure rates furthest away from the confidence intervals, are only adjusted to the boundary of the HCC grouping. To illustrate, using the example in Section 4.3 of an issuer with a low HCC group failure rate of 23.9 percent, that issuer's group adjustment factor would change from 19.1 percent under the current methodology to 0.1 percent under this option based on the 2017 benefit year national failure rate statistics. Specifically, under this example, the issuer's group adjustment factor would be the difference between the issuer's 23.9 percent low group failure rate and the upper bound confidence interval for the low HCC group is 23.8 percent (23.9 percent – 23.8 percent = 0.1 percent). Thus, this option could directly address the "payment cliff" and remove the extreme impact of small differences in HCC accuracy for issuers whose failure rates are near the edges of the confidence intervals.

At the same time, however, this option minimizes the impact of HHS-RADV adjustments on risk scores and risk adjustment transfers – including those outlier issuers with high error rates who are furthest away from the confidence intervals. As seen in Appendix B, in comparison to the current methodology, this option (the Confidence Intervals Methodology) would only adjust outlier issuers' risk scores at a fraction of the rate of the current methodology and would result in a significantly lower financial impact for all outlier issuers. For example, an issuer with a 70 percent failure rate in the high HCC group would be considered an outlier under the current methodology, having a failure rate more than 4 standard deviations away from the national mean, well beyond the 1.96 standard deviations required to be determined to have outlier status. A truly average issuer would have a 0.004 percent chance of having a failure rate this high due to random chance alone. As such, the example issuer is clearly an outlier and ought to receive an appropriate adjustment to its risk score(s) due to HHS-RADV. If adjusting to the mean, as under the current methodology, to bring this example issuer on par with the average issuer, the example issuer would receive a group adjustment factor of 70 percent – 26.2 percent = 43.8 percent. In comparison, if the issuer were adjusted to the edge of the confidence interval, they would receive a group adjustment factor of 70 percent – 47.1 percent = 22.9 percent, which reflects only a fraction of the misreported risk that negatively impacted the risk adjustment transfers of other issuers in the state market risk pool.

For this reason, we have concerns that this option would result in under-adjustments based on HHS-RADV results for issuers furthest away from the confidence intervals. As an extension of the findings in Appendix B, we are concerned this option results in such a minimal financial payment impact for outlier issuers that it may not deter up-coding in risk adjustment. Under this option, we found that the maximum and minimum error rates, reflecting the issuers who were furthest in their failure rates from the average issuer across all HCC failure rate groups are much lower in magnitude than under the current methodology and the sliding scale options described

later in this section. Specifically, the maximum error rate when adjusting all outliers to the confidence intervals would be around 15 percent, compared to a maximum error rate of 29.13 percent in 2017 benefit year HHS-RADV data for most of the other options under consideration. This suggests that under this option, even the most egregious failure rate outliers would receive minimal HHS-RADV adjustments. Therefore, although this option could address the "payment cliff" effect for issuers just outside of the confidence interval, this option may also create the unintended consequence of mitigating the financial impact for situations where issuers are not close to the confidence intervals.

4.4.3 Only Make Adjustments for Positive Error Rate Outliers

Another option suggested by some stakeholders that could address, at least in part, the "payment cliff" is to modify the current two-sided approach to HHS-RADV and only make adjustments for positive error rate outlier issuers. This option would retain the current calculation and associated thresholds to identify positive error rate outliers and would align with the policy finalized in the 2020 Payment Notice for exiting issuers.[84] This option may be attractive for non-outlier issuers because it could be seen as more predictable, as it limits the number of issuers whose risk scores would be adjusted as a result of HHS-RADV and would not result in adverse adjustments for zero error rate issuers based on negative error rate outliers.

We have concerns about only adjusting for positive error rate outliers for non-exiting issuers.[85] The intent of the two-sided outlier identification, and the resulting adjustments to outlier issuer risk scores that have significantly better-than-average or poorer-than-average data validation results is to ensure that HHS-RADV makes adjustments for identified, material risk differences between what issuers submitted to the EDGE servers and what was validated by the issuer's medical records. This ensures that, consistent with the statute, the HHS-operated risk adjustment program is transferring funds from issuers with plans with lower-than-average actuarial risk to issuers with plans with higher-than-average actuarial risk. Under this approach, HHS-RADV uses the two-sided outlier identification to ensure that the issuer who is coding well is able to recoup funds that might have been lost through HHS-RA because its competitors are coding badly. For example, if one issuer was fairly accurate in reporting their data to EDGE, resulting in a two percent HCC group failure rate in a state market risk pool, and another issuer had a tendency to report more conditions to EDGE than could be validated from the medical records, resulting in a twenty-five percent failure rate in the state market risk pool, in the absence of HHS-RADV, the issuer with the higher HCC group failure rate who had been reporting more conditions to EDGE than could be validated, would have unfairly benefited in risk adjustment (receiving a higher payment or lower charge amount), negatively impacting the issuer with the lower failure rate when considering the outcome of HHS-RA alone. Under a two-sided HHS-RADV risk adjustment, the lower failure rate issuer (as a negative outlier) would be able to recoup what would have been lost to the high failure rate issuer had data validation not been

[84] In the 2020 Payment Notice, we finalized a policy, applicable beginning with the 2018 benefit year of HHS-RADV, to only make adjustments to an exiting issuer's risk scores if it was determined to be a positive error rate outlier. See 84 FR at 17503-17504.

[85] See, e.g., the 2020 Payment Notice, 84 FR at 17504-17508.

performed. This logic mirrors how positive error rate outliers are treated whereby the issuer that had a higher than average number of validation errors is penalized for their higher number of errors and the rest of the issuers in the state market risk pool are able to recoup their losses for that higher than average failure rate issuer. Therefore, HHS-RADV uses a two-sided approach to, among other things, make adjustments to equalize the varying coding failure rates across issuers.[86]

However, some stakeholders have expressed concern that negative error rate outlier status may not be the result of issuers having fewer coding errors, but rather as a result of poor EDGE data submission. These stakeholders have suggested that adjusting for negative error rate outliers may reward issuers for submitting incomplete data to EDGE or reduce the incentive for issuers to submit accurate EDGE data. We do not agree that adjusting negative error rate outliers creates such an incentive, because the inherent risk of relying on receiving a negative error rate outlier status in HHS-RADV is too high to significantly interfere with incentives that issuers face to submit complete and accurate EDGE data. For example, we do not believe that there is an incentive for issuers, as a long-term strategy, to under-code in risk adjustment in hopes that enough of their under-coding is picked up in HHS-RADV for the issuer to be identified as a negative error rate outlier every year. We believe that negative error rate issuers that are under-coding will likely reassess their coding practices to ensure that they are accurately capturing their risk in future benefit years of risk adjustment through their initial EDGE submissions. We do not believe these issuers will want to wait for HHS-RADV to take place for a given benefit year in the hopes that they will be able to recoup any of these losses. In addition, as detailed elsewhere in this paper, we believe it is appropriate to include found HCCs to account for HCCs that were miscoded as another HCC within the same hierarchy and to ensure that charges are collected from issuers with lower-than-average actuarial risk and payments are made to issuers with higher-than-average actuarial risk.

We additionally believe that the potential to be a negative error rate outlier could incentivize all issuers to aim for the lowest possible failure rate instead of only aiming for a failure rate that is not a positive error rate outlier. Specifically, we believe that over time, this two-sided approach to adjusting risk scores for both positive and negative error rate outliers will put additional pressure on issuers to code more accurately.

In addition to suggestions to only make adjustments for positive error rate outliers, we have also received comments from stakeholders recommending that we consider treating negative error rate outliers differently than positive error rate outliers. For example, one suggestion was to calculate the group adjustment factor for the negative error rate outliers to the confidence interval and calculate the group adjustment factor for the positive error rate outliers to the weighted mean. Similar to the reasons outlined above in support of a two-sided outlier identification

[86] It is important to note the HHS-RADV approach is fundamentally different than the MA-RADV approach. MA-RADV only adjusts for positive error rate outliers, as the program's intent is to recoup federal funding that was the result of improper payments under the Medicare Part C program, which is not the intent of HHS-RADV.

process, we believe that positive and negative error rate outliers should generally be subject to the same process.

However, we recognize in the cases of negative error rate issuers with negative failure rates that calculating those issuers' group adjustment factor to the weighted mean may not be providing the appropriate incentives for issuers to code correctly in EDGE. Therefore, we discuss this issue in a later section of this chapter and explore an interim option to adjust those issuers' error rate calculations.

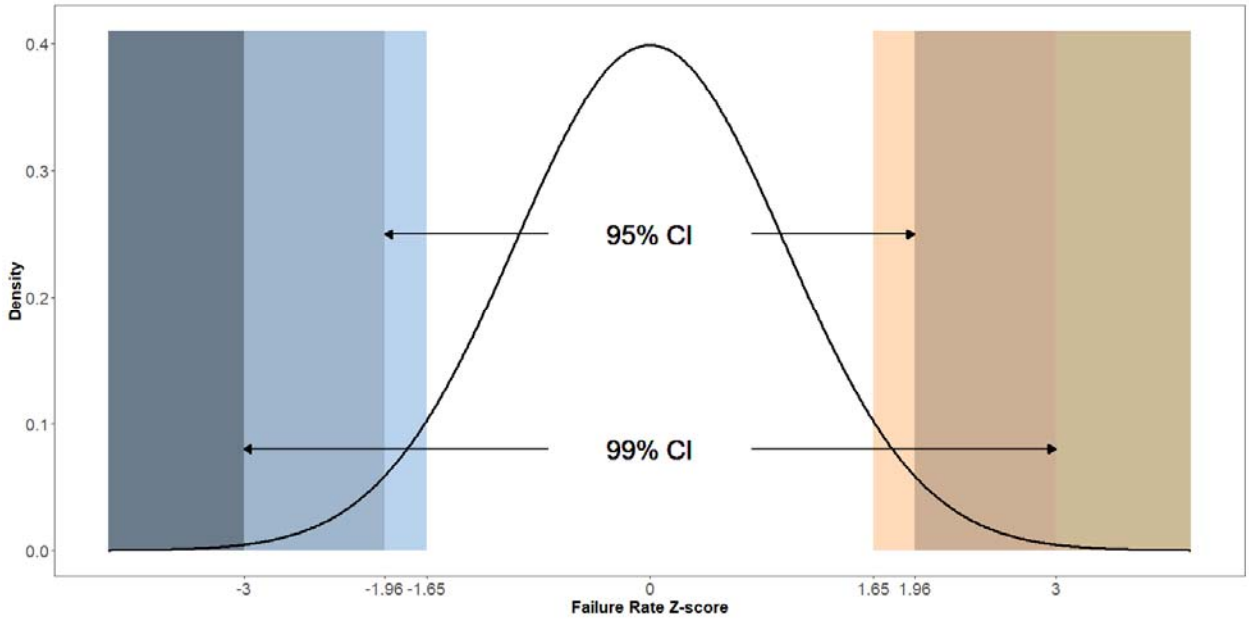
We further note that only adjusting for positive error rate outliers does not address the "payment cliff" and would retain the "payment cliff" for positive error rate outliers (see Appendix B). Therefore, for all of these reasons, we are concerned about moving to a one-sided outlier identification process and are instead interested in potentially pursuing other modifications to the current two-sided outlier identification process to address stakeholders' concerns.

4.4.4 Sliding Scale Adjustment Options

An alternative option to modify the calculation and application of outlier issuers' error rates to mitigate the impact of the "payment cliff" in cases where the failure rates are near the confidence interval is to calculate the group adjustment factor on a sliding scale basis. As discussed in the 2020 Payment Notice, we stated that we may consider alternative options for error rate adjustments, such as using multiple or smoothed confidence intervals for outlier identification and risk score adjustment.[87] If we were to pursue this option, we would need to select additional thresholds to create the sliding scale. Under the current methodology, using the standard normal distribution analogy in the figure below, risk scores for issuers that are inside the 95 percent confidence interval around the mean are not adjusted, and risk scores for issuers that lie outside of the 95 percent confidence interval are adjusted.

[87] 84 FR at 17507.

Figure 4.3: Normal distribution of confidence interval thresholds

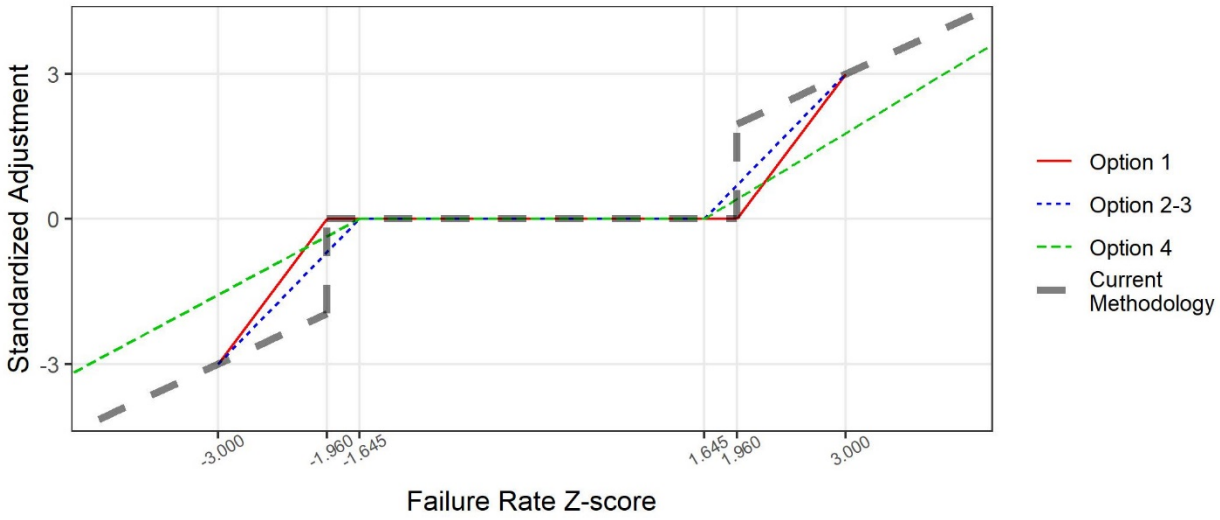


Applying a sliding scale adjustment to the error rate calculation could provide a more balanced approach to mitigate the "payment cliff" under the current methodology without potentially resulting in over-or under-adjusting issuers as a result of HHS-RADV, and would take into account the magnitude of the individual issuer's failure rate in applying an adjustment to the issuer's risk scores. Depending on the thresholds used under this option, it could also ensure that the approach to mitigate the "payment cliff" does not impact situations where outlier issuers' failure rates are not close to the confidence intervals.

Since finalizing the 2020 Payment Notice and calculating the 2017 benefit year HHS-RADV results, we have analyzed various options to explore the creation of a sliding scale adjustment to issuers' error rate calculations based on each issuer's distance from the confidence interval, using a variety of different thresholds. Figure 4.4 shows a comparison of the threshold options that we explored for purposes of developing this paper.

To apply this sliding scale adjustment, we used a linear formula that can be calculated using different threshold options. Under the options described in this section, issuers whose failure rates are near the point where the "payment cliff" occurs would be linearly adjusted between: (1) a failure rate value that occurs at the edge of a confidence interval; and (2) the group mean failure rate in the following form: $A = a\,FR + b$, where the coefficients a (the slope) and b (intercept) would be calculated based on the empirical HHS-RADV failure rate results for each HCC group (see Table 4.5). Using this linear sliding scale adjustment, all of these options could theoretically provide a smoothing effect in the error rate calculation for issuers just outside of the confidence interval. While we are exploring several sliding scale threshold options that are described in this white paper, we are also interested in comments on other potential thresholds for the sliding scale adjustment calculation.

Figure 4.4: Comparison of current error rate adjustment methodology with sliding scale adjustments Options 1-4 described in this section



As noted in Figure 4.4, Option 1 would create the sliding scale adjustment from +/- 1.96 to 3 standard deviations. This option would retain the confidence interval at 1.96 standard deviations under the current methodology, meaning that issuers within the 95 percent confidence interval would not have their respective risk scores adjusted. This option would also retain the full adjustment to the mean failure rate for issuers outside of the 99 percent confidence interval (beyond 3 standard deviations). The distinction of this option would be that it would adjust outlier issuers' error rates on a sliding scale between the 95 percent and 99 percent confidence interval bounds (1.96 to 3 standard deviations). This option retains the most aspects of the current methodology, which would provide stability for issuers. Option 1 also keeps the current significant adjustment to the HCC group weighted mean after 3 standard deviations to ensure the mitigation of the "payment cliff" for those close to the confidence intervals does not impact situations where outlier issuers' failure rates are not close to the confidence intervals.

Option 2, on other hand, would create a sliding scale adjustment from +/-1.645 to 3 standard deviations. This option would adjust the upper and lower bounds of the confidence interval to be at 1.645 standard deviations, meaning that issuers outside of the 90 percent confidence interval would have their risk scores adjusted, instead of beginning adjustments at the 95 percent confidence interval under the current methodology. This option would also adjust issuers' risk scores on a sliding scale between the 90 percent and 99 percent confidence intervals (between 1.645 and 3 standard deviations). This would mean that more issuers would be considered outliers under this option than the current methodology, as seen in Table 4.6 below.

Similar to Option 1, this option would retain the adjustment to the mean failure rate for issuers beyond the 99 percent confidence interval (outside 3 standard deviations). This option, in comparison to Option 1, could provide a more gradual smoothing effect for issuers just outside of the confidence interval, as seen in the bar chart below in Figure 4.7 where more issuers would have errors rates close to zero. However, even though this option lowers the overall impact of HHS-RADV adjustments to transfers, this option increases the number of outliers between 1.645

and 1.96 standard deviations and therefore, would increase the number of state market risk pools seeing adjustment to transfers as a result of HHS-RADV.

The third option that we are considering (Option 3) is to adjust risk scores for issuers that fall between +/-1.96 to 3 standard deviations, as in Option 1, but calculate the amount of the linear adjustment based on values between 1.645 and 3 standard deviations. Option 3 combines using the sliding scale adjustment values from Option 2, with retaining the current confidence intervals under Option 1. Specifically, this option would adjust issuers' risk scores on a sliding scale between 1.96 and 3 standard deviation as in Option 1 (between the 95 percent and 99 percent confidence interval bound) with a different magnitude for the linear adjustment for these issuers from Option 2. This means that this option retains the confidence intervals at 1.96 standard deviations, such that issuers within the 95 percent confidence interval would not have their risk scores adjusted, and it retains the adjustment to the mean failure rate for issuers beyond 3 standard deviations (outside of the 99 percent confidence interval).

Our theory was that Option 3 could provide the more gradual smoothing effect for issuers from Option 2 without increasing the number of issuers identified as outliers. However, this option could create a new, smaller "payment cliff" effect for issuers that lie outside of either side of the 1.96 threshold because the application of the linear adjustment factor would not apply until 1.96 standard deviations. This option also adds another layer of complexity to the error estimation methodology as it would include a set of issuers between 1.645 and 1.96 standard deviations in calculating the linear adjustment for the error rate, but exclude those issuers when applying that linear adjustment to issuers' error rate calculations.

The last option that we considered (Option 4) was to create a sliding scale adjustment starting +/-1.645 to the maximum failure rate z score. Option 4 would adjust the confidence intervals to start at 1.645 standard deviations, meaning that issuers outside the 90 percent confidence interval would have their risk scores adjusted (as in Option 2) and the linear adjustment would be applied until the maximum failure rate z score. Out of all of the options, this option would come the closest to eliminating any "payment cliff" in the error rate calculation for all issuers who are close to the confidence intervals. Because issuers beyond 3 standard deviations impact the calculation of the sliding scale adjustment factor under this option, Option 4 should have the least transfer impact on individual issuers that are just outside of the confidence intervals, as it has the lowest weighted mean of absolute transfer change over premiums. However, as seen in Table 4.6, this option increases the number of outliers between 1.645 and 1.96 standard deviations (as in Option 2), and as a result, an increased number of state market risk pools would have transfers adjusted as a result of HHS-RADV in comparison to the current methodology. Our concern with this option is that it may result in under-adjustments because even very extreme outliers would not receive the full adjustment back to the mean failure rate—the adjustment factor would be applied to failure rates beyond 3 standard deviations. Also, like Option 3, this option is more complex to calculate than Options 1 and 2 and is dependent on outliers that are not close to the confidence intervals.

4.4.5 Evaluating the Sliding Scale Adjustment Options

To assess the four sliding scale options described in the previous section, we ran a series of analyses using the 2017 benefit year HHS-RADV results to evaluate which option may best address the "payment cliff" issue to meet the stated policy goal.

First, to assist in comparing these sliding scale options, we compiled the slopes, number of outliers, and error rates for each of these options using the 2017 benefit year HHS-RADV results. Because Options 2 and 3 use the same values between 1.645 and 3 standard deviations, the slopes are the same in the below Table 4.5, even though there is a difference in the number of outliers identified under these options, as shown in the below Table 4.6. Options 2 and 4 have different slopes because they use a different end points of linear adjustments to calculate their slopes, but these options have the same increased number of identified outliers because these options would flag issuers as outliers starting at 1.645 standard deviations.[88]

Table 4.5: Comparing the Slopes for Sliding Scale Adjustment Options 1-4 described in this section

Method	HCC Group	Starting/End points of Linear Adjustment (z-scores)		Lower		Upper	
				a	b	a	b
		Lower	Upper	slope	intercept	slope	intercept
1	Low	-1.96 / -3	1.96 / 3	2.885	0.413	2.885	-0.687
	Medium	-1.96 / -3	1.96 / 3	2.885	0.114	2.885	-1.008
	High	-1.96 / -3	1.96 / 3	2.885	-0.155	2.885	-1.357
2 and 3	Low	-1.645 / -3	1.645 / 3	2.214	0.249	2.214	-0.46
	Medium	-1.645 / -3	1.645 / 3	2.214	0.018	2.214	-0.704
	High	-1.645 / -3	1.645 / 3	2.214	-0.193	2.214	-0.968
4	Low	-1.645 / -5.62	1.645 / 5.94	1.413	0.159	1.383	-0.287
	Medium	-1.645 / -16.69	1.645 / 5.64	1.109	0.009	1.412	-0.449
	High	-1.645 / -11.86	1.645 /6.94	1.161	-0.101	1.311	-0.573

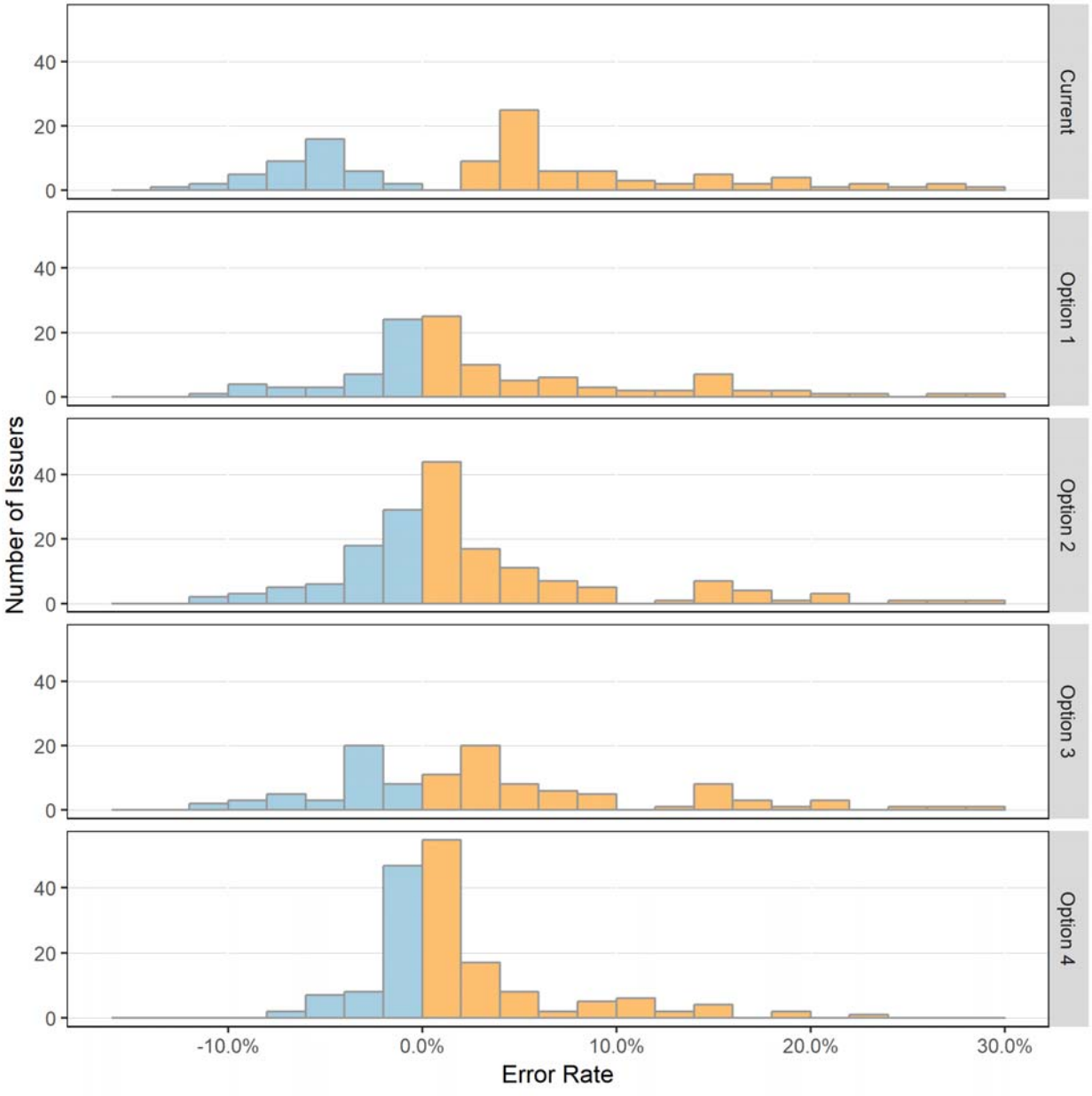
[88] We note that an outlier issuer can be a positive outlier in one HCC group and a negative outlier in another HCC group; therefore, this outlier issuer's error rate can change from being a positive error rate under one option to a negative error rate outlier under another option.

Table 4.6: Comparing Outlier Issuers Impacted by the Sliding Scale Adjustment Options

Method	# Positive Error Rate Issuers	# Negative Error Rate Issuers	Total Error Rate Issuers
Current Methodology	69	41	110
Option 1	68	42	110
Option 2	103	63	166
Option 3	69	41	110
Option 4	102	64	166

Next, we tested the differences in error rates between the current methodology and the various sliding scale options using the 2017 benefit year HHS-RADV results. As shown in Figure 4.7 below, we found that all of the sliding scale options under consideration in this white paper resulted in the distribution of error rates moving closer to zero than the current methodology. As expected, we found that Option 4, which had the smallest group adjustment factors for issuers including those beyond the 99 percent confidence interval, resulted in the smallest error rates. The maximum error rate for Option 4 was also smaller than any of the other sliding scale options being considered.

Figure 4.7: Comparing the Distribution of Estimated Error Rates between the Options Described in this section using the 2017 BY RADV Results



To further assess these options, we wanted to consider whether they directly mitigated the "payment cliff" for issuers just outside the confidence intervals. As previously mentioned, we expected that the more gradual slope from Option 2 that smooths the HCC group level adjustment factors could be better at mitigating the issuer level payment cliff than Option 1.

We did not find that to be the case in testing these sliding scale options using the 2017 benefit year HHS-RADV results. Because Options 1 and 2 use a different set of outliers, it was difficult to create a direct comparison of error rates. Under Option 1, the issuers just outside the confidence intervals were not the same issuers as those just outside the confidence intervals under Option 2. This meant that any comparison using issuers just outside of the confidence

intervals under Option 2 would always result in Options 1 and 3 having lower error rates overall because the risk scores of the issuers just outside the confidence interval in Option 2 were not being adjusted under Options 1 and 3.

For this reason, we compared the sliding scale options utilizing a moving-window approach by dynamically selecting a set of issuers in the evaluation of all sliding scale candidate options. The moving-window allowed us to set various boundaries of the z-score between (±)1.64 and (±)3.05, which covered the "issuers of interest" (i.e. those issuers that are just outside of the confidence intervals under all options to various degrees). The moving-window puts issuers that are closest based on their average failure rate z-scores into one group to evaluate the "payment cliff" via comparing the error rate difference of issuers within each group. Applying the moving-window method to the 2017 benefit year HHS-RADV data, we found that, compared to other options, Option 2 minimizes the "payment cliff" for issuers that are just outside or just inside the upper 95 percent confidence interval. However, for issuers that are close to the upper 90 percent confidence interval, or with lower-than-average failure rates among the issuers of interest (i.e., potential outliers), Option 1 outperforms Option 2 in reducing the "payment cliff".

These observations can be explained by the non-linear relationship between an issuer's error rate and the respective failure rate z-scores in the three HCC groups. The design of the sliding scale option is to smooth out the adjustment factors at the HCC group level. However, at the issuer level, the flattened adjustment factors due to a smaller slope in Option 2 compared to Option 1 could be diminished by other dominant factors. For example, at the HCC group level, Option 1 outperforms Option 2 because it results in zero adjustment factors and potentially a smaller "payment cliff" for issuers with failure rate z-scores between (±)1.645 and (±)1.96 in at least one HCC group. On the other hand, Option 2 outperforms Option 1 for issuers and HCC groups when failure rate z-scores that are between (±)1.96 and (±)3 because of the smaller slope and flattened adjustment factors in Option 2. Therefore, the overall performance of Options 1 and 2 in reducing the "payment cliff" at the issuer level depends on the number of issuers with failure rates z-scores between (±)1.645 and (±)1.96, compared to that of the issuers between (±)1.96 and (±)3 z-scores.

Lastly, to assist in comparing the sliding scale options to the other options described in this paper, in Appendix C, we provide the results of a simulation of the 2017 benefit year HHS-RADV in the same manner as Appendix B. These results show the estimated overall transfer and issuer impact of the sliding scale options on the 2017 benefit year HHS-RADV results. The estimated transfer impact of the sliding scale options between the current methodology and the four sliding scale options is generally as expected.

In short, all of the sliding scale options discussed in this paper result in lower error rates than the current methodology. The nonlinearity of error rates dilutes the ability to compare the effect of the four scaling scale options. Because all of the analyses comparing the impact of each "payment cliff" mitigation option are only based on 2017 benefit year HHS-RADV results (the first non-pilot year), issuers' future error rates may follow different patterns.

We are interested in comments from stakeholders regarding the options to mitigate the existing "payment cliff" for potential future rulemaking, outlined above, recognizing the current limitations related to having only year of non-pilot year data for testing purposes. In particular, we are interested in stakeholders' perspectives regarding what our priorities should be in considering these options, and which best create incentives and outcomes in line with the goals of the RA and RADV programs.

4.5 Negative Error Rate Issuers with Negative Failure Rates

As described earlier in this paper, the purpose of HHS-RADV is to promote confidence and stability in the budget-neutral transfer methodology used by the HHS-operated risk adjustment program by ensuring the integrity and quality of data provided by issuers. As described earlier in this chapter, one the purposes of the two-sided adjustment in HHS-RADV is to penalize issuers that validate HCCs in HHS-RADV at much lower rates than the national average and to reward issuers in HHS-RADV that validate HCCs in HHS-RADV at rates that are much higher than the national average, encouraging issuers to ensure that their EDGE-reported risk scores reflect the true actuarial risk of their enrollees. Positive and negative error outliers represent these two types of adjustments, respectively. An issuer can be identified as a negative error rate outlier due a number of contributing reasons; this section focuses on those issuers for whom low failure rates are driven by newly found HCCs rather than by high validation rates.

The current methodology does not distinguish between low failure rates due to accurate data submission and those that have been depressed through the presence of found HCCs. When a large number of found HCCs appear in an issuer's HHS-RADV sample, failure rates may be so low as to become negative. Although we are considering longer term options to more precisely identify true outliers and to address hierarchy considerations in HCC groups, an interim approach to mitigate the impact of HHS-RADV adjustments as a result of negative error rate outliers with negative failure rates would be to add a constraint in the group adjustment factor calculation in the current error rate calculation methodology for these issuers.

Specifically, we are considering constraining negative error rate outlier issuers' error rate calculation in cases when an issuer's failure rate is negative as a temporary measure. Currently, an outlier issuer's error rate is calculated based on the difference between the weighted mean failure rate for the HCC group and the issuer's failure rate for that HCC group, which may be a negative failure rate. We are considering adding a constraint to the group adjustment factor whereby negative failure rate issuers' error rates are calculated as the difference between the weighted mean failure rate for the HCC group (if positive) and zero (0). To illustrate, we would be substituting the following highlighted terms into the error rate process:

If $GFR_i^G > UB^G$ or $GFR_i^G < LB^G$:
Then $Flag_i^G = "outlier"$ and $Adjustment_i^G = GFR_{i,constr}^G - \mu(GFR^G)_{constr}$
If $GFR_i^G \leq UB^G$ and $GFR_i^G \geq LB^G$:
Then $Flag_i^G = "not\ outlier"$ and $Adjustment_i^G = 0$
Where:

GFR_i^G is an issuer's failure rate for the HCC failure rate grouping

$GFR_{i,constr}^G$ is an issuer's failure rate for the HCC failure rate grouping, constrained to 0 if GFR_i^G is less than 0. Also expressed as:

$$GFR_{i,constrained}^G = max\{0, GFR_i^G\}$$

$\mu(GFR^G)$ is the weighted national mean failure rate for the HCC failure rate grouping

$\mu(GFR^G)_{constr}$ is the weighted national mean failure rate for the HCC failure rate grouping, constrained to 0 if $\mu(GFR^G)$ is less than 0. Also expressed as:

$$\mu(GFR^G)_{constr} = max\{0, \mu(GFR^G)\}$$

UB^G and LB^G are the upper and lower bounds of the HCC failure rate grouping confidence interval, respectively.

$Flag_i^G$ is the indicator if issuer i's group failure rate for group G locates beyond a calculated threshold that we are using to classify issuers into "outliers" or "not outliers" for group G.

$Adjustment_i^G$ is the calculated adjustment amount to adjust issuer i's EDGE risk scores for all sampled HCCs in group G.

We would then compute total adjustments and risk adjustment transfer error rates for each issuer based on the sums of the $Adjustment_i^G$.[89]

This approach would limit the financial impact of adjustments due to negative error rate outliers with negative failure rates on other issuers, providing stability to issuers in predicting the HHS-RADV impact. For example, under the current error rate calculation using the 2017 benefit year HHS-RADV metrics, a negative outlier issuer with a -15 percent failure rate for the low HCC grouping would currently receive a group adjustment factor of the difference between -15 percent and the weighted mean for the low HCC grouping of 4.8 percent of -19.8 percent, but if we were to constrain the negative failure rates for negative outlier issuers to zero, the group adjustment factor in this example would be the difference between 0 percent and the weighted mean for the low HCC grouping of 4.8 percent, resulting in a -4.8 percent group adjustment factor. We believe that this type of constraint could help ensure that negative error rate issuers are rewarded for high validation rates while mitigating any incentive for under-reporting on EDGE.

We believe this option would not have a chilling effect on issuer data accuracy and could be easily implemented under the current methodology as a temporary stand-alone adjustment to the error rate calculation, or in combination with the previously discussed alternative options to calculate the error rate in this chapter. As described in Chapter 3 of this paper, we are considering options to account for HCCs miscoded into the same hierarchy and to address newly found HCCs that may be contributing to negative failure rates. These long-term changes could have an impact on the determination of the lower bound confidence interval and reduce the occurrence of negative failure rates, but would also represent substantial departures from the current error estimation methodology. Therefore, we believe that the addition of this type of

[89] See, for example, the 2018 Benefit Year Protocols: PPACA HHS Risk Adjustment Data Validation, Version 7.0 (June 24, 2019), available at https://www.regtap.info/reg_librarye.php?i=2904.

constraint to the current error rate calculation may offer a balanced interim option to mitigate the impact of negative outlier issuer adjustments on other issuers in the state market risk pool.

4.6 ALTERNATIVE OPTIONS

In addition to the aforementioned options, we received feedback on other potential changes to the calculation and application of issuers' error rates that we did take under consideration in drafting this white paper, but that were not specifically designed to mitigate the impact of the "payment cliff" or negative outlier issuer adjustments. For example, one recommendation was to subject error rate outliers to a second round of sampling and outlier determination before making an adjustment to risk scores and risk adjustment transfers. Our understanding is that the purpose of this alternative option would be to ensure those issuers identified as outliers are truly outliers by conducting a second round of auditing. We are concerned about the significant burden increase that this approach would create on issuers and HHS, as it would result in some issuers being required to conduct medical record retrieval and other IVA activities for two separate sets of enrollees for the same benefit year HHS-RADV. This type of approach would also delay when we would be able to provide issuers with HHS-RADV results.

In the past, we have also heard from stakeholders that the application and calculation of the error rate adjustment should take into consideration state differences in coding practices – that providers in some states may be better at coding than providers in other states, and that when HHS-RADV determines outlier status at the national level, the identification of outliers does not take those state-level differences into account. So far, we have not observed trends in the unmodified 2016 benefit year HHS-RADV results and the 2017 benefit year HHS-RADV results that indicate there is an overall significant difference among states' failure rate results compared to the national benchmarks, but we intend to continue to assess future HHS-RADV results to see if any trends in this regard emerge.

5. APPLICATION OF HHS-RADV RESULTS

This chapter considers a change to the application of HHS-RADV results to better reflect actuarial risk of the benefit year being audited. In the 2020 Payment Notice, we stated that while we are interested in applying the HHS-RADV results to the benefit year being audited, we have concerns about how to switch to that policy and adjust risk scores for a given benefit year twice.[90] This chapter considers options on how HHS might transition away from the current prospective application of HHS-RADV results[91] and move to an approach that would apply the results to the benefit year being audited.

5.1 OVERVIEW OF THE APPLICATION OF HHS-RADV RESULTS

In the 2014 Payment Notice, we finalized that HHS would use a prospective approach when making transfer adjustments based on findings from the data validation process.[92] Currently, HHS generally uses an issuer's HHS-RADV error rate from the prior year to adjust the issuer's average risk score in the current transfer year.[93] We finalized the use of a prospective approach to allow issuers and HHS sufficient time to complete the validation and appeals processes before transfer adjustments are made. As such, we generally used 2017 benefit year HHS-RADV results to adjust 2018 benefit year risk adjustment risk scores, resulting in an adjustment to 2018 benefit year risk adjustment transfer amounts.[94] In light of the policy finalized in the 2020 Payment Notice that delays collection, disbursement, and reporting of transfer adjustments to reflect HHS-RADV results[95], and the changes recently finalized to the risk adjustment holdback policy[96], we are considering whether to change this prospective approach to the application of HHS-RADV findings.

Specifically, we are considering applying HHS-RADV results to the same risk adjustment benefit year risk scores and transfers. For example, 2021 benefit year HHS-RADV results could be applied to adjust 2021 benefit year risk adjustment risk scores and transfers. Under this policy, the risk adjustment risk scores and transfers would only be adjusted based on the same benefit year's HHS-RADV results.[97]

We believe this change has the potential to provide stability for issuers and help them better predict the impact of HHS-RADV results. When we finalized the policy in the 2014 Payment Notice, we did not anticipate the extent of the changes that would occur in the risk profile of enrollees in the individual and small group markets from year to year or the changes in issuer market participation from year to year. Therefore, we believe that this potential change

---

[90] The exception to the current prospective application of HHS-RADV results is for exiting issuers, whose risk score error rates are applied to the PLRS and transfer amounts for the benefit year being audited.
[91] See 84 FR at 17507.
[92] See 78 FR 15410 at 15438.
[93] The exception to this general rule is for exiting issuers. See, supra note 90.
[94] Ibid.
[95] See 84 FR at 17506 – 17507.
[96] Available at: https://www.cms.gov/CCIIO/Resources/Regulations-and-Guidance/Downloads/Change-to-Risk-Adjustment-Holdback-Policy-for-the-2018-Benefit-Year-and-Beyond.pdf.
[97] Risk scores and risk adjustment transfer amounts may be subsequently adjusted in response to successful appeals.

could help address stakeholder concerns about maintaining actuarial soundness in the application of an issuer's HHS-RADV error rate if an issuer's risk profile, enrollment, or market participation changes substantially from year to year. We also believe that this type of change could eliminate the need to adjust each benefit year twice when there are issuers who have been identified as outliers exiting all of the market risk pools in a state (that is, not selling or offering any new plans in the state). It could also prevent cases where an issuer who enters a state market risk pool is subject to the adjustments for the HHS-RADV results from the prior benefit year when other issuers in the state market risk pool are outliers, even though those issuers did not participate in the state market risk pool for that HHS-RADV benefit year. For these reasons, we are interested in considering this potential change for future benefit years.

5.2 TRANSITION YEAR OPTIONS

Our main concern with implementing this option is the transition from the current prospective adjustment approach into an approach that would apply error rates to the benefit year being audited. In theory, if we were to implement this policy, we would apply two benefit years of HHS-RADV to one year of risk adjustment risk scores. For example, if we were to finalize and implement this policy for 2021 benefit year HHS-RADV, 2021 benefit year risk adjustment risk scores and transfers would be adjusted first to reflect 2020 benefit year HHS-RADV results, and then a second time based on 2021 benefit year HHS-RADV results.[98] Once implemented, for subsequent benefit years, risk adjustment risk scores and transfers would only be adjusted based on the same benefit year's HHS-RADV results.[99]

As we assess the options on how to move away from the prospective framework for future benefit years, we are specifically interested in comments on how we could approach the transition year and we are currently considering three options for how to do so.

First, if we implement this policy for 2021 benefit year HHS-RADV, one option (the average error rate option) would be to calculate an average value between 2021 and 2020 benefit years HHS-RADV error rates and apply this average error rate to 2021 risk adjustment risk scores and transfers. We believe this type of approach would be methodologically straightforward, and would help mitigate the potential impact of two HHS-RADV adjustments on a single year of risk adjustment risk scores, without adversely impacting the predictability of HHS-RADV on risk adjustment transfers. This option would combine the 2020 and 2021 HHS-RADV results into one set of results to be used to adjust 2021 benefit year risk adjustment risk scores and transfers; and therefore, this option would result in no separate RADV adjustment calculation for 2020 benefit year HHS-RADV results. However, as with the options mentioned below, this would result in one final adjustment amount to be collected and paid on the 2021 benefit year HHS-RADV timeline, in early 2025.

---

[98] In this illustrative example, it is possible that 2020 risk adjustment risk scores and transfers could be adjusted a third time in response to successful HHS-RADV appeals.

[99] See supra note 97.

Another option (the RA transfer option) would be to calculate 2020 benefit year HHS-RADV adjustments to 2021 benefit year risk adjustment transfers and 2021 benefit year HHS-RADV adjustments to 2021 benefit year risk adjustment transfers separately, then calculate the difference between each of these values and the unadjusted 2021 benefit year risk adjustment transfers before any benefit years HHS-RADV adjustments were applied, and add these differences together to arrive at the total HHS-RADV modification to the 2021 benefit year risk adjustment transfers. That is, HHS would calculate adjustments under 2020 and 2021 benefit years HHS-RADV and incorporate 2020 and 2021 benefit year HHS-RADV results applied to 2021 benefit year risk adjustment transfers in one final adjustment amount to be collected and paid on the 2021 benefit year HHS-RADV timeline, in early 2025.

A third option (the combined PLRS option) would be to apply 2020 benefit year HHS-RADV risk score adjustments to 2021 PLRSs, and then apply 2021 HHS-RADV risk score adjustments to the adjusted 2021 PLRSs. We would then use the final adjusted PLRSs (reflecting both the 2020 and 2021 HHS-RADV results) to adjust 2021 benefit year risk adjustment transfers. Like the RA transfer option, HHS would calculate adjustments under 2020 and 2021 benefit year HHS-RADV and incorporate 2020 and 2021 benefit year HHS-RADV results applied to 2021 benefit year risk adjustment transfers in one final adjustment amount to be collected and paid on the 2021 benefit year HHS-RADV timeline, in early 2025.

We are concerned that at least one of these options could result in duplication of the prior year's impacts for some issuers that had the same underlying issue for both years and therefore, we solicit comment on these options. We are specifically interested in comments on these alternative options to calculating HHS-RADV adjustments for a transition year that would move the program from a prospective application of these adjustments to applying HHS-RADV results to the same risk adjustment benefit year PLRS and transfers. We are also interested in comments on: (1) the advantages and disadvantages of any of these options; (2) which calculation option most closely aligns with the goals of the HHS-RADV program; and (3) whether we should be considering other options for the transition year.

6. CONCLUSION

After two pilot years, HHS has proceeded with making adjustments to reflect HHS-RADV results to ensure the integrity of the HHS-operated risk adjustment program by confirming that issuers can validate the risk that is being used to calculate risk adjustment transfers. The 2017 benefit year is the first non-pilot year where HHS-RADV results were used to adjust risk scores and risk adjustment transfers. The findings from the initial years of HHS-RADV indicate that most issuers' enrollee samples are representative and meet precision targets, that outlier detection issues are only occurring in limited cases where issuers have unusually low or high numbers of HCCs in an HCC group, and that the current methodology results in a more stable level of transfer changes based on HHS-RADV results than the original methodology.

As in all programs of this complexity, we recognize there are aspects that can be refined for future benefit years, such as the incorporation of measures to mitigate the impact of the "payment cliff" and transitioning to apply HHS-RADV results to the benefit year being audited. We look forward to feedback from stakeholders and the general public on the options presented in this paper and anticipate this feedback will inform the development of potential modifications to the HHS-RADV program for future benefit years. As noted in previous sections, the purpose of this paper is to seek stakeholder feedback at this time on the options that we are considering to address these policy issues prior to conducting rulemaking in these areas.

Commenters should submit comments by Monday, January 6, 2020 to CCIIOACARADataValidation@cms.hhs.gov with the subject line of "December 2019 HHS-RADV White Paper."

APPENDIX A: OVERVIEW OF HHS-RADV REGULATIONS

- March 11, 2013: HHS Notice of Benefit and Payment Parameters for 2014 (78 FR 15410) established the six steps of error estimation in § 153.630.
- March 11, 2014: HHS Notice of Benefit and Payment Parameters for 2015 (79 FR 13744):
    - Established the sample size, stratification and Neyman allocation;
    - Established IVA standards, SVA processes and that enrollee risk score validation would be based on medical record review;
    - Established error estimation process whereby issuers' plan enrollee average risk score is adjusted for any error, regardless of the size or magnitude of the error; and
    - Provided appeals, oversight, and data security standards.
- February 27, 2015: HHS Notice of Benefit and Payment Parameters for 2016 (80 FR 10750) increased the risk adjustment user fee to cover the administrative costs of HHS-RADV.
- December 22, 2016: HHS Notice of Benefit and Payment Parameters for 2018 (81 FR 94058):
    - Exempted issuers within the materiality threshold $15 million or less in premiums from participating in HHS-RADV except approximately every three years;
    - Required issuers to provide pharmacy claims to the IVA; and
    - Created a discrepancy reporting process for the audit sample, SVA results, and error rate calculation.
- April 17, 2018: HHS Notice of Benefit and Payment Parameters for 2019 (83 FR 16930):
    - Amended error estimation to only calculate and adjust issuers' risk scores when an issuer's failure rate is statistically significant based on three HCC groupings (low, medium, and high);
    - Exempted issuers with 500 or fewer billable member months from HHS-RADV;
    - Established that the IVA sample only includes enrollees from state risk pools with more than one issuer;
    - Permitted abbreviated mental health assessments in lieu of complete medical records when state privacy laws restrict the disclosure of mental health medical records; and
    - Clarified provisions regarding civil money penalties and adjustments due to demographic or enrollment errors discovered during HHS-RADV.
- April 25, 2019: HHS Notice of Benefit and Payment Parameters for 2020 (84 FR 17454):
    - Extended the Neyman allocation to the 10^{th} stratum for HHS-RADV sampling;
    - Clarified the application and distribution of default data validation charges;
    - Expanded the SVA to audit the full IVA sample when issuers failed pairwise means testing;
    - Adopted and piloted a methodology for including RXCs for the 2018 benefit year HHS-RADV;
    - Outlined the process for applying error rates for exiting issuers and sole issuer markets;
    - Updated the timeline for collection, distribution and reporting of HHS-RADV adjustments to transfers to provide more options to states and issuers for accounting for these amounts in rates and medical loss ratio reports; and
    - Codified HHS-RADV exemptions for issuers within the materiality thresholds (except approximately every three years), 500 or fewer billable member months, and in liquidation.

APPENDIX B: COMPARING THE 2017 BENEFIT YEAR HHS-RADV RESULTS USING THE CURRENT ERROR RATE METHODOLOGY, ORIGINAL ERROR RATE METHODOLOGY, CONFIDENCE INTERVALS METHODOLOGY, AND ONLY POSITIVE METHODOLOGY IN CHAPTER 4[100]

Individual Market Risk Pools – 2018 Risk Adjustment				
Metrics	Current Methodology	Original Methodology	Confidence Intervals Methodology	Only Positive Error Rate Outlier Methodology
Total Risk Adjustment Transfers before RADV	$4,008,083,759	$4,008,083,759	$4,008,083,759	$4,008,083,759
Total Risk Adjustment Transfers after RADV	$4,018,098,320	$3,883,342,860	$4,016,365,468	$3,986,049,393
Total RADV Payment Transfer Amounts	$329,819,454	$2,018,305,677	$49,235,794	$150,981,462
Total RADV Charge Transfer Amounts	-$329,819,454	-$2,018,305,677	-$49,235,794	-$150,981,462
Percent RADV Payment Transfers Over Total Transfers Before RADV	8.23%	50.36%	1.23%	3.77%
Issuer's Average Absolute Transfer over Premium	0.89%	5.27%	0.13%	0.41%
Member Weighted Risk Score	1.547	1.547	1.547	1.547
Member Weighted Risk Score with RADV	1.553	1.448	1.549	1.542
Risk Score % Change	0.35%	-6.87%	0.10%	-0.33%
% Billable Member Months by issuers with Adjusted Risk Scores	15.3%	70.5%	15.3%	2.5%
# State Market Risk Pool	51	51	51	51
# State Market Risk Pools with RADV Adjustments	18	44	18	8
# Issuers	258	258	258	258
# Issuers with Adjusted Risk Scores	28	190	28	10
# Issuers with Adjusted RA Transfers	127	237	127	73
# Issuers with Reduced Transfers After RADV	87	113	78	10
# Issuers with Increased Transfers After RADV	40	124	49	63
% of Issuers with Adjusted RA Transfers	49.2%	91.9%	49.2%	28.3%

[100] Catastrophic risk pools were excluded from the results for the individual market. Results for merged market states (Massachusetts and Vermont) are reported as part of the individual market. Because 2017 benefit year HHS-RADV was a pilot year for Massachusetts, Massachusetts issuers' results are counted in the before RADV and after RADV payments totals, but those issuers have zero error rates under all options; therefore, the state market risk pool is not adjusted in the 2017 HHS-RADV results for all options in this tables.

Small Group Market Risk Pools – 2018 Risk Adjustment				
Metrics	Current Methodology	Original Methodology	Only Adjusting to Confidence Intervals Methodology	Only Positive Methodology
Total Risk Adjustment Transfers before RADV	$1,161,924,456	$1,161,924,456	$1,161,924,456	$1,161,924,456
Total Risk Adjustment Transfers after RADV	$1,226,212,243	$1,464,926,038	$1,155,673,750	$1,253,776,026
Total RADV Payment Transfer Amounts	$346,330,506	$1,407,927,984	$58,040,017	$122,709,965
Total RADV Charge Transfer Amounts	-$346,330,506	-$1,407,927,984	-$58,040,017	-$122,709,965
Percent RADV Payment Transfers Over Total Transfers Before RADV	29.81%	121.17%	5.00%	10.56%
Issuer's Average Absolute Transfer over Premium	1.26%	5.39%	0.21%	0.40%
Member Weighted Risk Score	1.270	1.270	1.270	1.270
Member Weighted Risk Score with RADV	1.279	1.176	1.272	1.265
Risk Score % Change	0.68%	-8.01%	0.17%	-0.39%
% Billable Member Months by issuers with Adjusted Risk Scores	22.1%	86.2%	22.1%	3.4%
# State Market Risk Pools	49	49	49	49
# State Market Risk Pools with RADV Adjustments	31	49	31	24
# Issuers	473	473	473	473
# Issuers with Adjusted Risk Scores	78	379	78	45
# Issuers with Adjusted RA Transfers	329	473	329	273
# Issuers with Reduced Transfers After RADV	207	247	214	45
# Issuers with Increased Transfers After RADV	122	226	115	228
% of Issuers with Adjusted RA Transfers	69.6%	100.0%	69.6%	57.7%

Individual Market Risk Pools – 2017 Risk Adjustment				
Metrics	Current Methodology	Original Methodology	Confidence Intervals Methodology	Only Positive Methodology
Total Risk Adjustment Transfers before RADV	$3,870,537,132	$3,870,537,132	$3,870,537,132	$3,870,537,132
Total Risk Adjustment Transfers after RADV	$3,877,649,989	$3,928,448,874	$3,871,177,444	$3,871,598,886
Total RADV Payment Transfer Amounts	$21,194,560	$167,040,082	$3,945,316	$11,238,538
Total RADV Charge Transfer Amounts	-$21,194,560	-$167,040,082	-$3,945,316	-$11,238,538
Percent RADV Payment Transfers Over Total Transfers Before RADV	0.55%	4.32%	0.10%	0.29%
Issuer's Average Absolute Transfer over Premium	0.06%	0.41%	0.01%	0.03%
Member Weighted Risk Score	1.541	1.541	1.541	1.541
Member Weighted Risk Score with RADV	1.542	1.537	1.541	1.541
Risk Score % Change	0.00%	-0.26%	0.00%	-0.01%
% Billable Member Months by issuers with Adjusted Risk Scores	0.4%	2.1%	0.4%	0.1%
# State Market Risk Pools	51	51	51	51
# State Market Risk Pools with RADV Adjustments	15	28	15	13
# Issuers	391	391	391	391
# Issuers with Adjusted Risk Scores	18	42	18	16
# Issuers with Adjusted RA Transfers	161	279	160	135
# Issuers with Reduced Transfers After RADV	40	53	40	16
# Issuers with Increased Transfers After RADV	121	226	120	119
% of Issuers with Adjusted RA Transfers	41.2%	71.4%	40.9%	34.5%

Small Group Market Risk Pools – 2017 Risk Adjustment[101]				
Metrics	Current Methodology	Original Methodology	Only Adjusting to Confidence Intervals Methodology	Only Positive Methodology
Total Risk Adjustment Transfers before RADV	$1,265,821,729	$1,265,821,729	$1,265,821,729	$1,265,821,729
Total Risk Adjustment Transfers after RADV	$1,266,388,710	$1,368,654,185	$1,265,927,261	$1,266,308,028
Total RADV Payment Transfer Amounts	$3,548,056	$173,053,167	$239,643	$993,404
Total RADV Charge Transfer Amounts	-$3,548,056	-$173,053,167	-$239,643	-$993,404
Percent RADV Payment Transfers Over Total Transfers Before RADV	0.28%	13.67%	0.02%	0.08%
Issuer's Average Absolute Transfer over Premium	0.01%	0.55%	0.00%	0.00%
Member Weighted Risk Score	1.271	1.271	1.271	1.271
Member Weighted Risk Score with RADV	1.271	1.267	1.271	1.271
Risk Score % Change	0.00%	-0.32%	0.00%	0.00%
% Billable Member Months by issuers with Adjusted Risk Scores	0.10%	2.00%	0.10%	0.00%
# State Market Risk Pools	48	48	48	48
# State Market Risk Pools with RADV Adjustments	7	28	7	4
# Issuers	498	498	498	498
# Issuers with Adjusted Risk Scores	9	44	9	6
# Issuers with Adjusted RA Transfers	113	331	113	67
# Issuers with Reduced Transfers After RADV	49	69	49	6
# Issuers with Increased Transfers After RADV	64	262	64	61
% of Issuers with Adjusted RA Transfers	22.7%	66.5%	22.7%	13.5%

[101] The 2017 benefit year small group market for the state of Ohio was excluded in this summary because there were manual adjustments to HHS-RADV transfer adjustments to correct for an issuer data submission discrepancy reflected in the Summary Report of 2017 HHS-RADV Adjustments to Transfers released on August 1, 2019, available at: https://www.cms.gov/CCIIO/Programs-and-Initiatives/Premium-Stabilization-Programs/Downloads/BY2017-HHSRADV-Adjustments-to-RA-Transfers-Summary-Report.pdf. For simulation purposes in Appendix B and C, excluding the state market risk pool with the manual adjustment allows the analysis to only reflect the impact due to the performance of the error estimation methods under consideration.

APPENDIX C: COMPARING THE 2017 BENEFIT YEAR HHS-RADV RESULTS USING SLIDING SCALE OPTIONS IN CHAPTER 4[102]

Individual Market Risk Pools – 2018 Risk Adjustment				
Metrics	Option 1	Option 2	Option 3	Option 4
Total Risk Adjustment Transfers before RADV	$4,008,083,759	$4,008,083,759	$4,008,083,759	$4,008,083,759
Total Risk Adjustment Transfers after RADV	$4,036,363,976	$4,043,719,836	$4,029,878,362	$4,030,247,409
Total RADV Payment Transfer Amounts	$136,966,244	$231,943,351	$202,308,778	$137,750,104
Total RADV Charge Transfer Amounts	-$136,966,244	-$231,943,351	-$202,308,778	-$137,750,104
Percent RADV Payment Transfers Over Total Transfers Before RADV	3.42%	5.79%	5.05%	3.44%
Issuer's Average Absolute Transfer over Premium	0.37%	0.61%	0.54%	0.36%
Member Weighted Risk Score	1.547	1.547	1.547	1.547
Member Weighted Risk Score with RADV	1.551	1.553	1.552	1.551
Risk Score % Change	0.25%	0.35%	0.28%	0.22%
% Billable Member Months by issuers with Adjusted Risk Scores	15.3%	24.6%	15.3%	24.6%
# State Market Risk Pool	51	51	51	51
# State Market Risk Pools with RADV Adjustments	18	29	18	29
# Issuers	258	258	258	258
# Issuers with Adjusted Risk Scores	28	51	28	51
# Issuers with Adjusted RA Transfers	127	186	127	186
# Issuers with Reduced Transfers After RADV	78	114	88	114
# Issuers with Increased Transfers After RADV	49	72	39	72
% of Issuers with Adjusted RA Transfers	49.2%	72.1%	49.2%	72.1%

[102] See supra note 100.

Small Group Market Risk Pools – 2018 Risk Adjustment				
Metrics	Option 1	Option 2	Option 3	Option 4
Total Risk Adjustment Transfers before RADV	$1,161,924,456	$1,161,924,456	$1,161,924,456	$1,161,924,456
Total Risk Adjustment Transfers after RADV	$1,181,600,625	$1,197,286,697	$1,194,592,346	$1,174,095,392
Total RADV Payment Transfer Amounts	$160,912,306	$246,180,867	$225,706,403	$144,761,922
Total RADV Charge Transfer Amounts	-$160,912,306	-$246,180,867	-$225,706,403	-$144,761,922
Percent RADV Payment Transfers Over Total Transfers Before RADV	13.85%	21.19%	19.43%	12.46%
Issuer's Average Absolute Transfer over Premium	0.58%	0.90%	0.82%	0.53%
Member Weighted Risk Score	1.270	1.270	1.270	1.270
Member Weighted Risk Score with RADV	1.276	1.277	1.277	1.274
Risk Score % Change	0.47%	0.57%	0.54%	0.35%
% Billable Member Months by issuers with Adjusted Risk Scores	22.1%	27.6%	22.1%	27.6%
# State Market Risk Pools	49	49	49	49
# State Market Risk Pools with RADV Adjustments	31	40	31	40
# Issuers	473	473	473	473
# Issuers with Adjusted Risk Scores	78	120	78	120
# Issuers with Adjusted RA Transfers	329	430	329	430
# Issuers with Reduced Transfers After RADV	214	273	215	273
# Issuers with Increased Transfers After RADV	115	157	114	157
% of Issuers with Adjusted RA Transfers	69.6%	90.9%	69.6%	90.9%

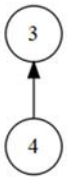
Individual Market Risk Pools – 2017 Risk Adjustment				
Metrics	Option 1	Option 2	Option 3	Option 4
Total Risk Adjustment Transfers before RADV	$3,870,537,132	$3,870,537,132	$3,870,537,132	$3,870,537,132
Total Risk Adjustment Transfers after RADV	$3,872,147,015	$3,873,913,097	$3,873,718,609	$3,872,048,467
Total RADV Payment Transfer Amounts	$9,679,664	$15,008,694	$13,791,460	$9,529,453
Total RADV Charge Transfer Amounts	-$9,679,664	-$15,008,694	-$13,791,460	-$9,529,453
Percent RADV Payment Transfers Over Total Transfers Before RADV	0.25%	0.39%	0.36%	0.25%
Issuer's Average Absolute Transfer over Premium	0.03%	0.04%	0.04%	0.03%
Member Weighted Risk Score	1.541	1.541	1.541	1.541
Member Weighted Risk Score with RADV	1.541	1.541	1.541	1.541
Risk Score % Change	-0.01%	-0.01%	0.00%	0.00%
% Billable Member Months by issuers with Adjusted Risk Scores	0.4%	0.4%	0.4%	0.4%
# State Market Risk Pools	51	51	51	51
# State Market Risk Pools with RADV Adjustments	15	19	15	19
# Issuers	391	391	391	391
# Issuers with Adjusted Risk Scores	18	23	18	23
# Issuers with Adjusted RA Transfers	161	204	161	204
# Issuers with Reduced Transfers After RADV	40	57	40	57
# Issuers with Increased Transfers After RADV	121	147	121	147
% of Issuers with Adjusted RA Transfers	41.2%	52.2%	41.2%	52.2%

Small Group Market Risk Pools – 2017 Risk Adjustment[103]				
Metrics	Option 1	Option 2	Option 3	Option 4
Total Risk Adjustment Transfers before RADV	$1,265,821,729	$1,265,821,729	$1,265,821,729	$1,265,821,729
Total Risk Adjustment Transfers after RADV	$1,266,149,552	$1,266,256,135	$1,266,234,531	$1,266,074,744
Total RADV Payment Transfer Amounts	$656,874	$1,979,327	$1,686,327	$1,125,448
Total RADV Charge Transfer Amounts	-$656,874	-$1,979,327	-$1,686,327	-$1,125,448
Percent RADV Payment Transfers Over Total Transfers Before RADV	0.05%	0.16%	0.13%	0.09%
Issuer's Average Absolute Transfer over Premium	0.00%	0.01%	0.01%	0.00%
Member Weighted Risk Score	1.271	1.271	1.271	1.271
Member Weighted Risk Score with RADV	1.271	1.271	1.271	1.271
Risk Score % Change	0.00%	0.00%	0.00%	0.00%
% Billable Member Months by issuers with Adjusted Risk Scores	0.1%	0.2%	0.1%	0.2%
# State Market Risk Pools	48	48	48	48
# State Market Risk Pools with RADV Adjustments	7	12	7	12
# Issuers	498	498	498	498
# Issuers with Adjusted Risk Scores	9	15	9	15
# Issuers with Adjusted RA Transfers	113	158	113	158
# Issuers with Reduced Transfers After RADV	49	64	49	64
# Issuers with Increased Transfers After RADV	64	94	64	94
% of Issuers with Adjusted RA Transfers	22.7%	31.7%	22.7%	31.7%

[103] See supra note 101.

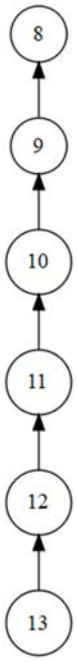
APPENDIX D: DIAGRAMS AND TABLES OF CURRENT HCC HIERARCHY STRUCTURE

Central Nervous System Infections



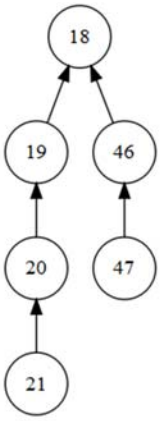
HCC	HCC Label
3	Central Nervous System Infections, Except Viral Meningitis
4	Viral or Unspecified Meningitis

Cancer



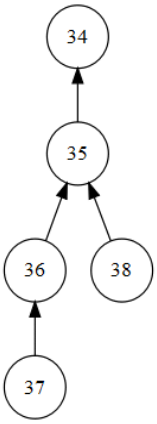
HCC	HCC Label
8	Metastatic Cancer
9	Lung, Brain, and Other Severe Cancers, Including Pediatric Acute Lymphoid Leukemia
10	Non-Hodgkin's Lymphomas and Other Cancers and Tumors
11	Colorectal, Breast (Age < 50), Kidney, and Other Cancers
12	Breast (Age 50+) and Prostate Cancer, Benign/Uncertain Brain Tumors, and Other Cancers and Tumors
13	Thyroid Cancer, Melanoma, Neurofibromatosis, and Other Cancers and Tumors

Pancreas Disorders



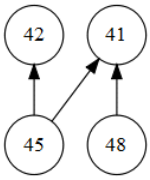
HCC	HCC Label
18	Pancreas Transplant Status/Complications
19	Diabetes with Acute Complications
20	Diabetes with Chronic Complications
46	Chronic Pancreatitis
47	Acute Pancreatitis/Other Pancreatic Disorders and Intestinal Malabsorption
21	Diabetes without Complication

Liver Disorders



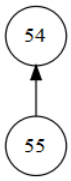
HCC	HCC Label
34	Liver Transplant Status/Complications
35	End-Stage Liver Disease
36	Cirrhosis of Liver
37	Chronic Hepatitis
38	Acute Liver Failure/Disease, Including Neonatal Hepatitis

Gastrointestinal Disorders



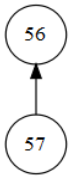
HCC	HCC Label
41	Intestine Transplant Status/Complications
42	Peritonitis/Gastrointestinal Perforation/Necrotizing Enterocolitis
45	Intestinal Obstruction
48	Inflammatory Bowel Disease

Necrosis



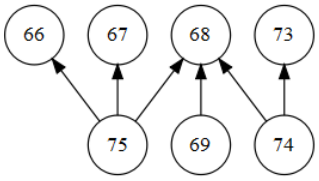
HCC	HCC Label
54	Necrotizing Fasciitis
55	Bone/Joint/Muscle Infections/Necrosis

Autoimmune Disorders



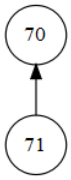
HCC	HCC Label
56	Rheumatoid Arthritis and Specified Autoimmune Disorders
57	Systemic Lupus Erythematosus and Other Autoimmune Disorders

Blood and Immune Disorders



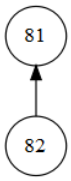
HCC	HCC Label
66	Hemophilia
67	Myelodysplastic Syndromes and Myelofibrosis
68	Aplastic Anemia
69	Acquired Hemolytic Anemia, Including Hemolytic Disease of Newborn
73	Combined and Other Severe Immunodeficiencies
74	Disorders of the Immune Mechanism
75	Coagulation Defects and Other Specified Hematological Disorders

Hemoglobin Disorders



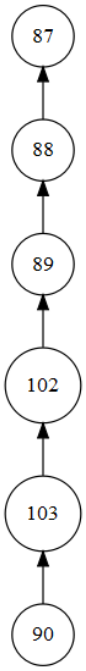
HCC	HCC Label
70	Sickle Cell Anemia (Hb-SS)
71	Thalassemia Major

Substance Use



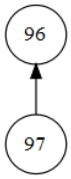
HCC	HCC Label
81	Drug Psychosis
82	Drug Dependence

Behavioral and Developmental Disorders



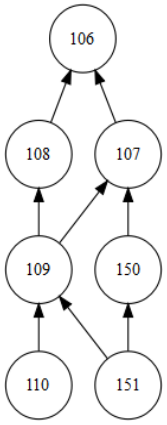
HCC	HCC Label
87	Schizophrenia
88	Major Depressive and Bipolar Disorders
89	Reactive and Unspecified Psychosis, Delusional Disorders
90	Personality Disorders
102	Autistic Disorder
103	Pervasive Developmental Disorders, Except Autistic Disorder

Chromosomal Syndromes



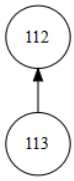
HCC	HCC Label
96	Prader-Willi, Patau, Edwards, and Autosomal Deletion Syndromes
97	Down Syndrome, Fragile X, Other Chromosomal Anomalies, and Congenital Malformation Syndromes

Paralysis



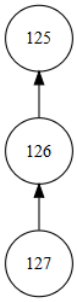
HCC	HCC Label
106	Traumatic Complete Lesion Cervical Spinal Cord
107	Quadriplegia
108	Traumatic Complete Lesion Dorsal Spinal Cord
109	Paraplegia
110	Spinal Cord Disorders/Injuries
150	Hemiplegia/Hemiparesis
151	Monoplegia, Other Paralytic Syndromes

Cerebral Palsy



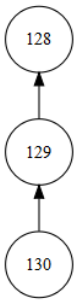
HCC	HCC Label
112	Quadriplegic Cerebral Palsy
113	Cerebral Palsy, Except Quadriplegic

Respiratory Distress



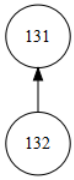
HCC	HCC Label
125	Respirator Dependence/Tracheostomy Status
126	Respiratory Arrest
127	Cardio-Respiratory Failure and Shock, Including Respiratory Distress Syndromes

Heart Failure



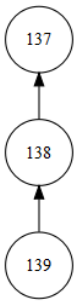
HCC	HCC Label
128	Heart Assistive Device/Artificial Heart
129	Heart Transplant
130	Congestive Heart Failure

Heart Disease



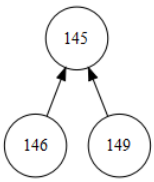
HCC	HCC Label
131	Acute Myocardial Infarction
132	Unstable Angina and Other Acute Ischemic Heart Disease

Heart Defects



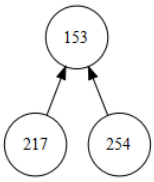
HCC	HCC Label
137	Hypoplastic Left Heart Syndrome and Other Severe Congenital Heart Disorders
138	Major Congenital Heart/Circulatory Disorders
139	Atrial and Ventricular Septal Defects, Patent Ductus Arteriosus, and Other Congenital Heart/Circulatory Disorders

Stroke



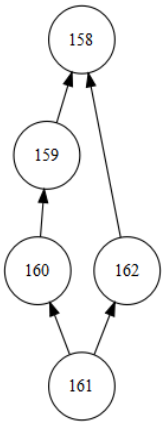
HCC	HCC Label
145	Intracranial Hemorrhage
146	Ischemic or Unspecified Stroke
149	Cerebral Aneurysm and Arteriovenous Malformation

Skin Ulcers and Amputation



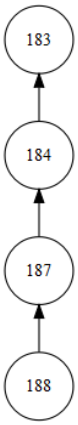
HCC	HCC Label
153	Atherosclerosis of the Extremities with Ulceration or Gangrene
217	Chronic Ulcer of Skin, Except Pressure
254	Amputation Status, Lower Limb/Amputation Complications

Pulmonary Disorders



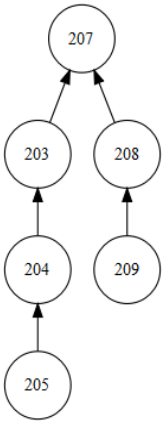
HCC	HCC Label
158	Lung Transplant Status/Complications
159	Cystic Fibrosis
160	Chronic Obstructive Pulmonary Disease, Including Bronchiectasis
161	Asthma
162	Fibrosis of Lung and Other Lung Disorders

Kidney Disease



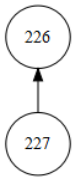
HCC	HCC Label
183	Kidney Transplant Status
184	End Stage Renal Disease
187	Chronic Kidney Disease, Stage 5
188	Chronic Kidney Disease, Severe (Stage 4)

Pregnancy



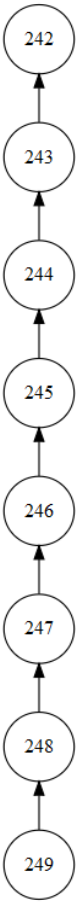
HCC	HCC Label
203	Ectopic and Molar Pregnancy, Except with Renal Failure, Shock, or Embolism
204	Miscarriage with Complications
205	Miscarriage with No or Minor Complications
207	Completed Pregnancy With Major Complications
208	Completed Pregnancy With Complications
209	Completed Pregnancy with No or Minor Complications

Fractures



HCC	HCC Label
226	Hip Fractures and Pathological Vertebral or Humerus Fractures
227	Pathological Fractures, Except of Vertebrae, Hip, or Humerus

Newborns



HCC	HCC Label
242	Extremely Immature Newborns, Birthweight < 500 Grams
243	Extremely Immature Newborns, Including Birthweight 500-749 Grams
244	Extremely Immature Newborns, Including Birthweight 750-999 Grams
245	Premature Newborns, Including Birthweight 1000-1499 Grams
246	Premature Newborns, Including Birthweight 1500-1999 Grams
247	Premature Newborns, Including Birthweight 2000-2499 Grams
248	Other Premature, Low Birthweight, Malnourished, or Multiple Birth Newborns
249	Term or Post-Term Singleton Newborn, Normal or High Birthweight

HCCs without a Hierarchy	
HCC	**HCC Label**
1	HIV/AIDS
2	Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock
6	Opportunistic Infections
23	Protein-Calorie Malnutrition
26	Mucopolysaccharidosis
27	Lipidoses and Glycogenosis
28	Congenital Metabolic Disorders, Not Elsewhere Classified
29	Amyloidosis, Porphyria, and Other Metabolic Disorders
30	Adrenal, Pituitary, and Other Significant Endocrine Disorders
61	Osteogenesis Imperfecta and Other Osteodystrophies
62	Congenital/Developmental Skeletal and Connective Tissue Disorders
63	Cleft Lip/Cleft Palate
64	Major Congenital Anomalies of Diaphragm, Abdominal Wall, and Esophagus, Age < 2
94	Anorexia/Bulimia Nervosa
111	Amyotrophic Lateral Sclerosis and Other Anterior Horn Cell Disease
114	Spina Bifida and Other Brain/Spinal/Nervous System Congenital Anomalies
115	Myasthenia Gravis/Myoneural Disorders and Guillain-Barre Syndrome/Inflammatory and Toxic Neuropathy
117	Muscular Dystrophy
118	Multiple Sclerosis
119	Parkinson's, Huntington's, and Spinocerebellar Disease, and Other Neurodegenerative Disorders
120	Seizure Disorders and Convulsions
121	Hydrocephalus
122	Non-Traumatic Coma, Brain Compression/Anoxic Damage
135	Heart Infection/Inflammation, Except Rheumatic
142	Specified Heart Arrhythmias
154	Vascular Disease with Complications
156	Pulmonary Embolism and Deep Vein Thrombosis
163	Aspiration and Specified Bacterial Pneumonias and Other Severe Lung Infections
251	Stem Cell, Including Bone Marrow, Transplant Status/Complications
253	Artificial Openings for Feeding or Elimination

APPENDIX E: TABLE OF HCC FAILURE RATE GROUPINGS FOR 2017 BENEFIT YEAR HHS-RADV

HCC	HCC Group	HCC Label
1	Low HCC Group	HIV/AIDS
2	Medium HCC Group	Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock
3	High HCC Group	Central Nervous System Infections, Except Viral Meningitis
4	High HCC Group	Viral or Unspecified Meningitis
6	High HCC Group	Opportunistic Infections
8	Medium HCC Group	Metastatic Cancer
9	High HCC Group	Lung, Brain, and Other Severe Cancers, Including Pediatric Acute Lymphoid Leukemia
10	Medium HCC Group	Non-Hodgkin's Lymphomas and Other Cancers and Tumors
11	High HCC Group	Colorectal, Breast (Age < 50), Kidney, and Other Cancers
12	High HCC Group	Breast (Age 50+) and Prostate Cancer, Benign/Uncertain Brain Tumors, and Other Cancers and Tumors
13	High HCC Group	Thyroid Cancer, Melanoma, Neurofibromatosis, and Other Cancers and Tumors
18	Low HCC Group	Pancreas Transplant Status/Complications
19	High HCC Group	Diabetes with Acute Complications
20	Low HCC Group	Diabetes with Chronic Complications
21	Low HCC Group	Diabetes without Complication
23	Medium HCC Group	Protein-Calorie Malnutrition
26	High HCC Group	Mucopolysaccharidosis
27	High HCC Group	Lipidoses and Glycogenosis
28	Medium HCC Group	Congenital Metabolic Disorders, Not Elsewhere Classified
29	High HCC Group	Amyloidosis, Porphyria, and Other Metabolic Disorders
30	Medium HCC Group	Adrenal, Pituitary, and Other Significant Endocrine Disorders
34	Medium HCC Group	Liver Transplant Status/Complications
35	Medium HCC Group	End-Stage Liver Disease
36	Low HCC Group	Cirrhosis of Liver
37	Medium HCC Group	Chronic Hepatitis
38	Medium HCC Group	Acute Liver Failure/Disease, Including Neonatal Hepatitis
41	Low HCC Group	Intestine Transplant Status/Complications
42	High HCC Group	Peritonitis/Gastrointestinal Perforation/Necrotizing Enterocolitis
45	High HCC Group	Intestinal Obstruction
46	Medium HCC Group	Chronic Pancreatitis
47	Medium HCC Group	Acute Pancreatitis/Other Pancreatic Disorders and Intestinal Malabsorption

HCC	HCC Group	HCC Label
48	Low HCC Group	Inflammatory Bowel Disease
54	High HCC Group	Necrotizing Fasciitis
55	Medium HCC Group	Bone/Joint/Muscle Infections/Necrosis
56	Low HCC Group	Rheumatoid Arthritis and Specified Autoimmune Disorders
57	Low HCC Group	Systemic Lupus Erythematosus and Other Autoimmune Disorders
61	High HCC Group	Osteogenesis Imperfecta and Other Osteodystrophies
62	Medium HCC Group	Congenital/Developmental Skeletal and Connective Tissue Disorders
63	High HCC Group	Cleft Lip/Cleft Palate
64	High HCC Group	Major Congenital Anomalies of Diaphragm, Abdominal Wall, and Esophagus, Age < 2
66	Medium HCC Group	Hemophilia
67	High HCC Group	Myelodysplastic Syndromes and Myelofibrosis
68	High HCC Group	Aplastic Anemia
69	High HCC Group	Acquired Hemolytic Anemia, Including Hemolytic Disease of Newborn
70	Medium HCC Group	Sickle Cell Anemia (Hb-SS)
71	Medium HCC Group	Thalassemia Major
73	High HCC Group	Combined and Other Severe Immunodeficiencies
74	High HCC Group	Disorders of the Immune Mechanism
75	Medium HCC Group	Coagulation Defects and Other Specified Hematological Disorders
81	High HCC Group	Drug Psychosis
82	High HCC Group	Drug Dependence
87	Low HCC Group	Schizophrenia
88	High HCC Group	Major Depressive and Bipolar Disorders
89	High HCC Group	Reactive and Unspecified Psychosis, Delusional Disorders
90	High HCC Group	Personality Disorders
94	Medium HCC Group	Anorexia/Bulimia Nervosa
96	Low HCC Group	Prader-Willi, Patau, Edwards, and Autosomal Deletion Syndromes
97	High HCC Group	Down Syndrome, Fragile X, Other Chromosomal Anomalies, and Congenital Malformation Syndromes
102	Low HCC Group	Autistic Disorder
103	Low HCC Group	Pervasive Developmental Disorders, Except Autistic Disorder
106	High HCC Group	Traumatic Complete Lesion Cervical Spinal Cord
107	High HCC Group	Quadriplegia
108	Medium HCC Group	Traumatic Complete Lesion Dorsal Spinal Cord
109	Low HCC Group	Paraplegia
110	High HCC Group	Spinal Cord Disorders/Injuries

HCC	HCC Group	HCC Label
111	High HCC Group	Amyotrophic Lateral Sclerosis and Other Anterior Horn Cell Disease
112	Low HCC Group	Quadriplegic Cerebral Palsy
113	Medium HCC Group	Cerebral Palsy, Except Quadriplegic
114	Low HCC Group	Spina Bifida and Other Brain/Spinal/Nervous System Congenital Anomalies
115	Medium HCC Group	Myasthenia Gravis/Myoneural Disorders and Guillain-Barre Syndrome/Inflammatory and Toxic Neuropathy
117	Low HCC Group	Muscular Dystrophy
118	Low HCC Group	Multiple Sclerosis
119	Medium HCC Group	Parkinson's, Huntington's, and Spinocerebellar Disease, and Other Neurodegenerative Disorders
120	Low HCC Group	Seizure Disorders and Convulsions
121	Medium HCC Group	Hydrocephalus
122	High HCC Group	Non-Traumatic Coma, Brain Compression/Anoxic Damage
125	Low HCC Group	Respirator Dependence/Tracheostomy Status
126	High HCC Group	Respiratory Arrest
127	High HCC Group	Cardio-Respiratory Failure and Shock, Including Respiratory Distress Syndromes
128	Low HCC Group	Heart Assistive Device/Artificial Heart
129	Medium HCC Group	Heart Transplant
130	Medium HCC Group	Congestive Heart Failure
131	High HCC Group	Acute Myocardial Infarction
132	High HCC Group	Unstable Angina and Other Acute Ischemic Heart Disease
135	High HCC Group	Heart Infection/Inflammation, Except Rheumatic
137	High HCC Group	Hypoplastic Left Heart Syndrome and Other Severe Congenital Heart Disorders
138	High HCC Group	Major Congenital Heart/Circulatory Disorders
139	High HCC Group	Atrial and Ventricular Septal Defects, Patent Ductus Arteriosus, and Other Congenital Heart/Circulatory Disorders
142	Medium HCC Group	Specified Heart Arrhythmias
145	High HCC Group	Intracranial Hemorrhage
146	High HCC Group	Ischemic or Unspecified Stroke
149	Medium HCC Group	Cerebral Aneurysm and Arteriovenous Malformation
150	Low HCC Group	Hemiplegia/Hemiparesis
151	High HCC Group	Monoplegia, Other Paralytic Syndromes
153	High HCC Group	Atherosclerosis of the Extremities with Ulceration or Gangrene
154	High HCC Group	Vascular Disease with Complications

HCC	HCC Group	HCC Label
156	High HCC Group	Pulmonary Embolism and Deep Vein Thrombosis
158	High HCC Group	Lung Transplant Status/Complications
159	Medium HCC Group	Cystic Fibrosis
160	Low HCC Group	Chronic Obstructive Pulmonary Disease, Including Bronchiectasis
161	Low HCC Group	Asthma
162	Medium HCC Group	Fibrosis of Lung and Other Lung Disorders
163	High HCC Group	Aspiration and Specified Bacterial Pneumonias and Other Severe Lung Infections
183	Low HCC Group	Kidney Transplant Status
184	High HCC Group	End Stage Renal Disease
187	Low HCC Group	Chronic Kidney Disease, Stage 5
188	Low HCC Group	Chronic Kidney Disease, Severe (Stage 4)
203	Low HCC Group	Ectopic and Molar Pregnancy, Except with Renal Failure, Shock, or Embolism
204	High HCC Group	Miscarriage with Complications
205	High HCC Group	Miscarriage with No or Minor Complications
207	High HCC Group	Completed Pregnancy With Major Complications
208	High HCC Group	Completed Pregnancy With Complications
209	Medium HCC Group	Completed Pregnancy with No or Minor Complications
217	Low HCC Group	Chronic Ulcer of Skin, Except Pressure
226	High HCC Group	Hip Fractures and Pathological Vertebral or Humerus Fractures
227	High HCC Group	Pathological Fractures, Except of Vertebrae, Hip, or Humerus
242	High HCC Group	Extremely Immature Newborns, Birthweight < 500 Grams
243	Medium HCC Group	Extremely Immature Newborns, Including Birthweight 500-749 Grams
244	Medium HCC Group	Extremely Immature Newborns, Including Birthweight 750-999 Grams
245	Medium HCC Group	Premature Newborns, Including Birthweight 1000-1499 Grams
246	High HCC Group	Premature Newborns, Including Birthweight 1500-1999 Grams
247	Low HCC Group	Premature Newborns, Including Birthweight 2000-2499 Grams
248	Medium HCC Group	Other Premature, Low Birthweight, Malnourished, or Multiple Birth Newborns
249	High HCC Group	Term or Post-Term Singleton Newborn, Normal or High Birthweight
251	Low HCC Group	Stem Cell, Including Bone Marrow, Transplant Status/Complications
253	Low HCC Group	Artificial Openings for Feeding or Elimination
254	Low HCC Group	Amputation Status, Lower Limb/Amputation Complications