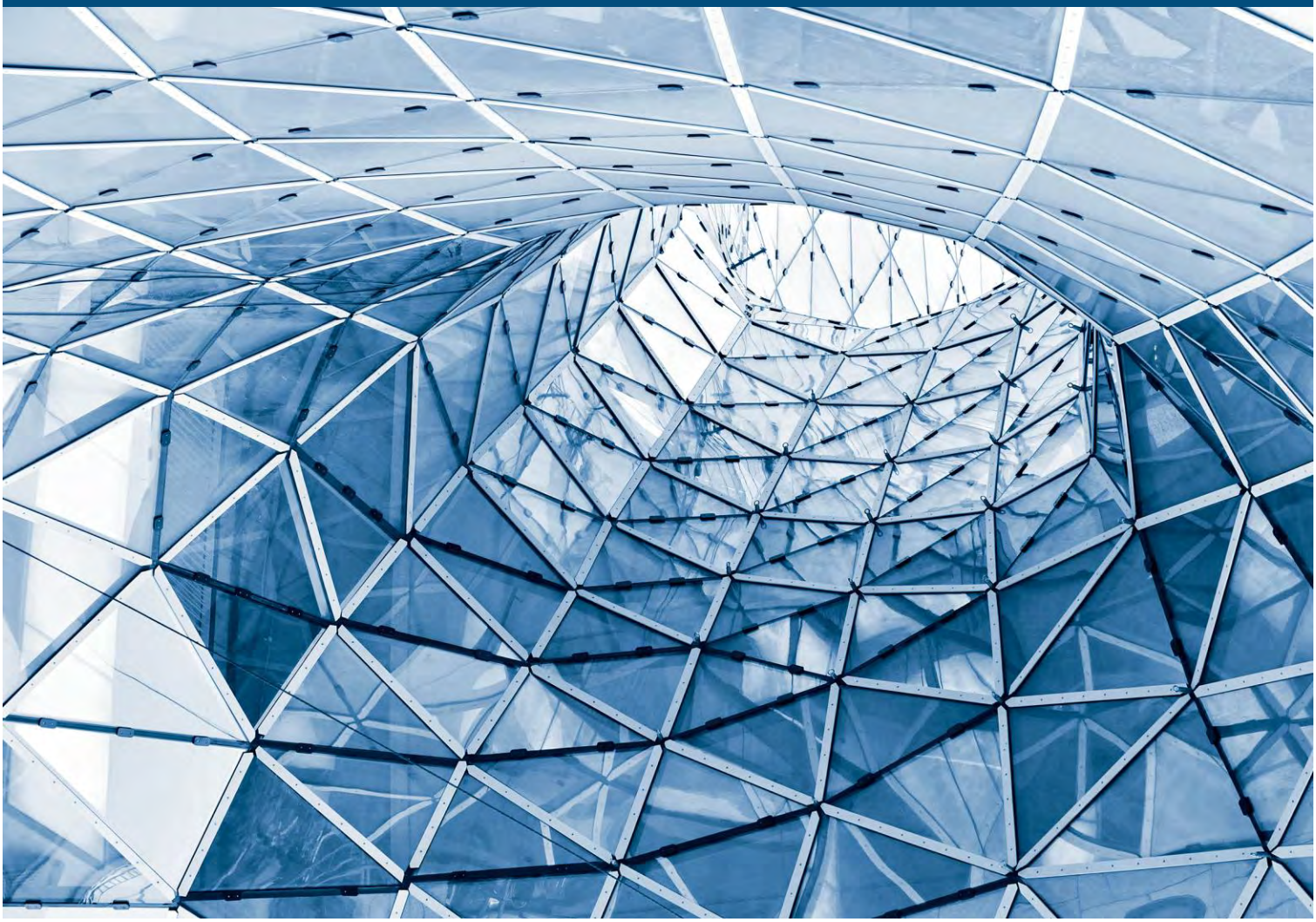
SOCIETY OF ACTUARIES

A Case Study of Risk Adjustment for Texas Medicaid



February 2017

A Case Study of Risk Adjustment for Texas Medicaid

SPONSOR Health Section

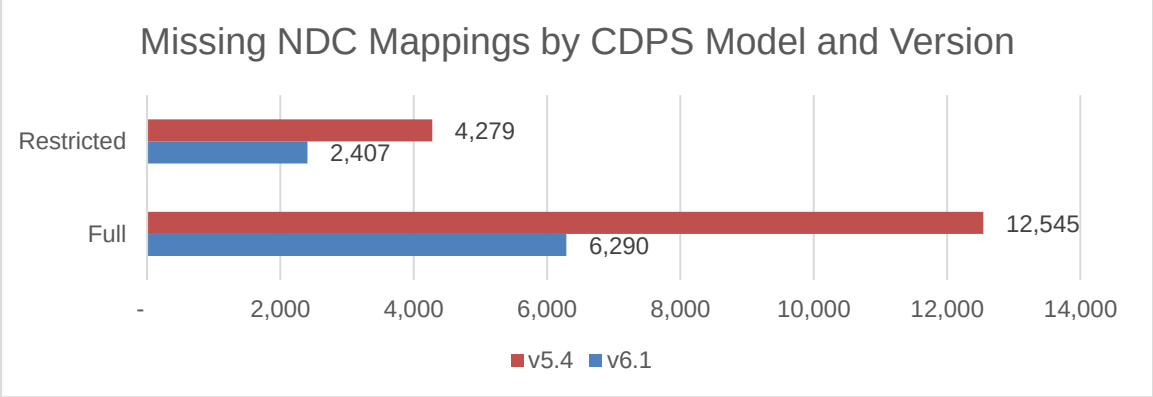
AUTHOR Brandon Barber, MS

Table of Contents

Executive Summary

The combination of frequent drug releases and less frequent risk factor mapping updates in risk adjustment models produces an undesired consequence where risk scores are deficient to intended values. These effects are less observable in models including medical claims since diagnostic data often identifies the same risk factors as pharmacy data and diagnosis revisions occur far less often. However, in capitated payment models using case mix ratios to risk adjustment premiums, minor deviations in risk scores can result in significant financial consequences to managed care organizations.

In a mapping of formulary equivalent drugs amongst existing risk factors in the Chronic Illness & Disability Payment System (CDPS) risk model, we identified the following number of missing national drug codes by model version. The latest available CDPS version (v6.1) demonstrates a more recent update to pharmacy mappings, but for our case study of Texas Medicaid, we continue to use version 5.4 since it is currently used for the most recent state fiscal year.



Separate from pharmacy mapping omissions and our recommendation for review organizations and actuaries to apply more regular updates independent of vendor model revisions, we identified unintended consequences due to recalibration. Our case study demonstrated two primary calibration concerns concerning ancillary spend exclusion when modeling total spend and voidance of clinical hierarchies. Regarding ancillary spend, models commonly exclude diagnoses from risk scoring for non-face-to-face encounters, but do not exclude their expenditures when modeling total spend. When these expenditures are excluded, it makes the assumption that every risk category carries ancillary spend proportional to their risk factor weight. However, our beneficiary sample demonstrated a contradictory result when isolating on members only within each risk factor category. Conditions such as cancer and pregnancy demonstrated vastly disproportional needs to other risk categories and MCOs with disproportionate share of such beneficiaries will exhibit unintended financial gains or losses as a result.

Our final calibration issue, voidance of clinical hierarchies, occurs when models no longer require clinical weight ranking orders as originally intended. The result produces events such as “extra high” infectious disease conditions such as AIDS carrying less weight than “medium” severity infectious diseases. The consequence means that risk models will suggest lower severity conditions require more expenditures than higher severity conditions.

Acknowledgements

We wish to thank the Society of Actuaries (SOA) Health Section Research Committee for sponsoring and Steven Siegel, SOA Research actuary, for his oversight of this study. We are also very appreciative of Rebecca Owen, SOA Health Researcher, for encouraging and facilitating various phases of the study. Geof Hileman of Kennell & Associates provided the essential initial impetus for our study while conducting his own research, *Accuracy of Claims-Based Risk Scoring Models* (2016), for the SOA Health Section Research Committee.

We would like to thank the following individuals from the Texas Health and Human Services Commission as well for providing extremely valuable feedback and information over the course of our study: Frank Genco, David Hufstedler, and Rachel Butler. Finally, we wish to thank Todd Gilmer, developer of CDPS, for providing vital information regarding model derivation and recalibration practices.

Geof Hileman, FSA, MAAA, Tina Liu, CPA, MBA, Marjorie Rosenberg, FSA, Sara Teppema, FSA, MAAA, and Steve Tutewohl, FSA, MAAA provided review of this report.

Purpose

Health risk adjustment models are an important component of premium revenue for many types of health plans, including Medicare Advantage, individual and small group plans post-Affordable Care Act, and Medicaid managed care plans. According to the Kaiser Commission on Medicaid and the Uninsured, 13 states use the Chronic Illness & Disability Payment System (CDPS) model and four states use the Medicaid Rx model, which are both developed by University of California, San Diego (UCSD) for Medicaid risk adjustment. CDPS is a diagnostic and/or pharmacy classification system that Medicaid programs use to make health-based capitated payments for beneficiaries (Gilmer, 2016). These algorithms are used to risk adjust capitation rates paid by the states to managed care organizations (MCOs).

Risk adjustment for MCOs is intended to measure expected cost variance across health plan member groups by evaluating the underlying health status of each population. In the state of Texas, for example, premium rates are risk adjusted based on the case mix ratio (defined below) of the plan-specific acuity score to the service area (grouped by geography) value within each defined risk group serviced by the MCOs. Texas MCOs define risk groups using age groupings for children while breaking adults into separate categories for Pregnant Women and Temporary Assistance for Needy Families (TANF).

$$Case\ Mix\ Ratio = \frac{Plan\ Predicted\ Expenditures\ Per\ Member\ Per\ Month}{Group\ Predicted\ Expenditures\ Per\ Member\ Per\ Month}$$

In state fiscal year 2015, risk adjustment transfers led to large financial impacts for Texas Medicaid MCOs, from losses of 42% of risk-based group premium revenue to gains of 32%. Thus, for states using CDPS in similar applications via rate setting, misapplication can result in significant financial consequences.

During a research study sponsored by the Society of Actuaries (SOA) to revisit the accuracy of claims based risk scoring models, led by Geof Hileman at Kennell and Associates Inc., the researchers identified a potential area for concern regarding risk scores derived from pharmacy data. Fit statistics for the pharmacy-only model of CDPS declined sharply in the Hileman study compared to a 2007 study (Winkelman and Mehmud, 2007) that other algorithms of similar design, also present in the 2007 study, did not exhibit. A review of the CDPS SAS programming code revealed the algorithm's mapping of National Drug Code

(NDC) to risk factors last occurred in early 2013 and according to a report by Express Scripts (Express Scripts, 2013), over $70 billion or approximately 20% of total 2013 drug spend, in drug patent expirations were expected during the four-year period from 2013-2016. Patent expirations shift drug utilization from brand names to new generic drugs and therefore to new NDC codes not present in the existing CDPS methodology. Due to this flurry of generic drugs in recent years, we hypothesized that the algorithm's pharmacy mappings inadvertently exclude new drugs adopted by prescribing physicians.

This study quantifies the impact of risk score deflation due to unintentional drug mapping exclusions and further explores additional issues that may arise in the application of risk scoring algorithms by performing a case study of Texas Medicaid MCOs.

Rate Setting Consequences

Accurately calculated risk scores are critical to MCOs and other risk-bearing entities using case mix ratios to adjust premium payments. We identified two major contributors to risk factor inaccuracy: outdated risk factor mappings due to new drugs and model calibration issues. Inaccuracy, due to either issue, directly affects issuer financial performance. While we lack the complete data repository necessary to fully quantify the rate setting impact for Texas Medicaid MCOs, we believe the data suggests a strong possibility exists for such inaccuracies and subsequent unintended financial consequences to the MCOs.

Unintentional drug-to-CDPS risk factor omissions exceeded 13% and 18% of current mappings for the full (pharmacy only) and restricted (diagnosis and pharmacy) CDPS model versions at the end of 2015 prior to the recent 6.1 release, respectively. Omissions were highly concentrated within very few individual risk factors and were most prevalent in older enrollment segments. The effect of NDC exclusions was mitigated by the simultaneous presence of diagnoses in the restricted CDPS model while the pharmacy-only (full) version demonstrated a much larger disparity between current and updated risk scores using generic equivalence mappings to identify unintentionally excluded drugs. At the plan and risk group level, where rate setting occurs for Texas MCOs, unintended risk score deficiency varied greatly. Furthermore, the issue is worsening with time; we found that over time, risk scores are becoming less accurate.

Consequences due to calibration issues were two-fold: (1) excluding certain items such as ancillary expenses from the dependent expenditure variable in model calibration introduces a potential bias and (2) insignificant terms or broken clinical hierarchies create model confusion and potential over-fitting issues that may not be suitable for new populations. The dependent calibration issue represents the greatest threat to case mix inaccuracies since it makes a proportional ancillary resource expenditure assumption that does not hold true via inspection of mutually exclusive enrollment segments. Our analysis identified two risk factor groupings, cancer and pregnancy, which demonstrated extreme deviance from population level ancillary spend while significant spend variance was observed across nearly all groupings. As we will demonstrate with plan level prevalence rates, combining disproportionate share of select morbidities with disproportionate ancillary service spend yields a highly likely scenario for improper risk factor weightings.

Study Design

Derivation Data

We derived data from four Texas MCOs participating in each of the last three state fiscal years (2013-2015), which consisted of nearly 15 million member months and $2.5 billion in allowed claims. The Texas Medicaid managed care program includes the Children's Health Insurance Program (CHIP) population and

the Temporary Assistance for Needy Families, referred to as STAR in Texas. Enrollees were included from any CHIP or STAR plan, who met either of the following conditions:

1. Enrolled continuously for at least six months with no more than a one month lapse in enrollment during the six month period
2. All infants less than one year of age and enrolled for at least one month

The state of Texas manages an additional Medicaid Managed Care Program called Star+Plus; our study did not include any members from this program.

Eligibility was further restricted based on state-defined risk groups for each program. STAR included enrollees from the following risk groups: Less than 1 Year of Age, Age 1 to 5, Age 6 to 14, Age 15 to 18, Age 19 to 20, Pregnant Women, and TANF Adults. CHIP enrollees were separately restricted to the following risk groups: Less than 1 Year of Age, Age 1 to 5, Age 6 to 14, and Age 15 to 18. Our resulting sample was distributed as follows in Table 1:

Table 1 – Risk Group and Program Distribution for Sampled Enrollees

Program	Risk Group	% of Total Sample
CHIP	Less than 1 Year of Age	< 1%
	Age 1 to 5	2.7%
	Age 6 to 14	9.5%
	Age 15 to 18	3.4%
STAR	Less than 1 Year of Age	4.1%
	Age 1 to 5	27.6%
	Age 6 to 14	39.0%
	Age 15 to 18	10.4%
	Age 19 to 20	< 1%
	Pregnant Women	1.5%
	TANF Adults	1.7%

Risk Scoring Application

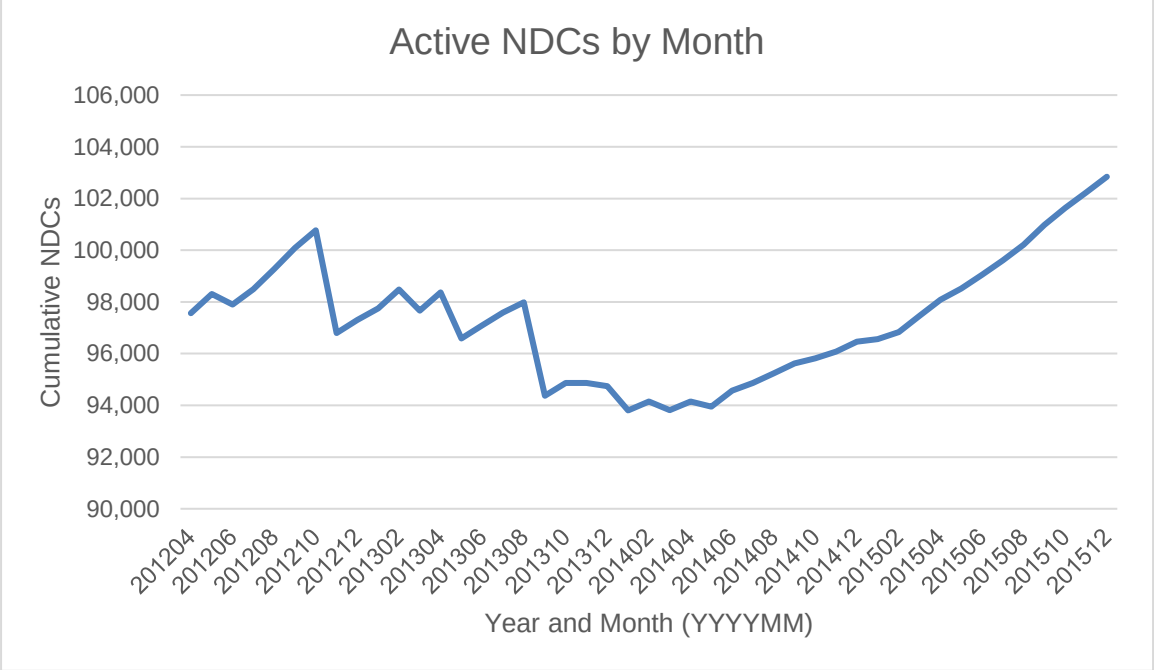
We followed the Texas Rate Analysis Technical Specifications for state fiscal year (SFY) 2015 provided by the Texas External Quality Review Organization (EQRO) in the application of risk adjustment. Our data was sourced from the three most recent state fiscal years (SFY13-SFY15) from four distinct Texas managed care organizations (MCOs) with enrollment in each SFY. Each of the selected MCOs provided consent to participate in our study.

As directed by state rating methodology, we applied the Chronic Illness and Disability Payment System (CDPS) version 5.4 SAS algorithm and associated concurrent weights relevant to each program SFY and enrollee type. The state of Texas applies the combined diagnosis and pharmacy model (Dx+Rx) along with state calibrated weights, which we have applied based on the reporting period for SFY15. In addition to the combined model, we applied the pharmacy-only model to our data sample to make inferences for states using the pharmacy-only model for capitation purposes. Since the state of Texas does not use this model for rate setting and we do not have the entire state database to accurately recalibrate, we used the base pharmacy-only CDPS version 5.4 concurrent weights in our risk adjustment application.

Pharmacy-Based Model Updates

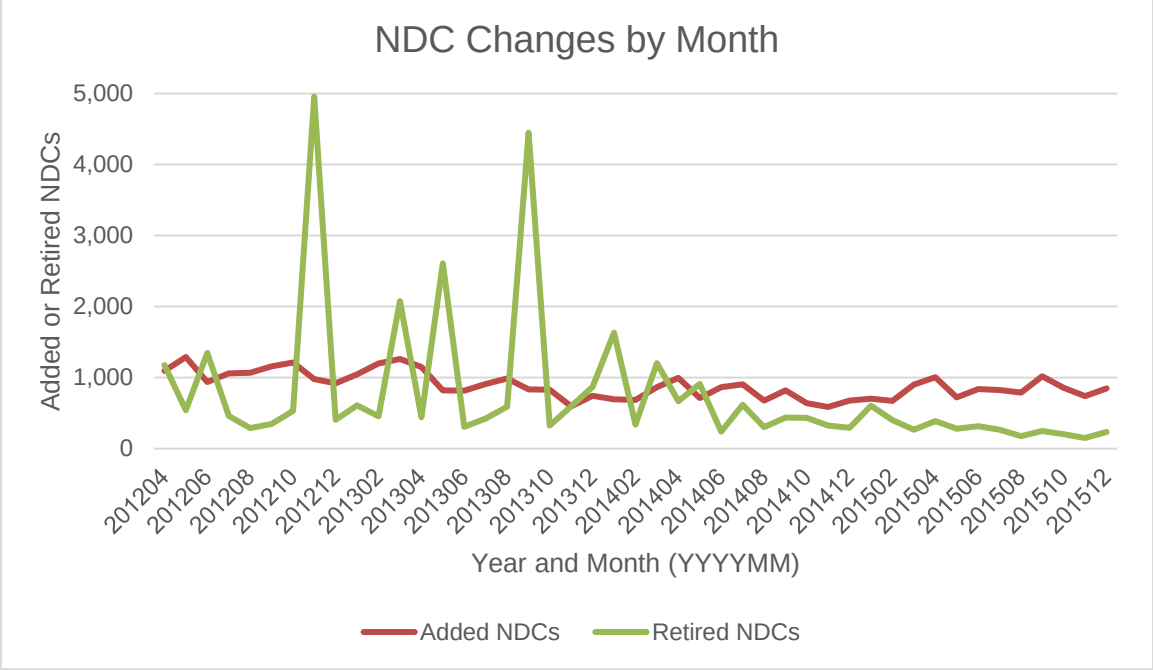
Claims-based model risk factors are primarily attributed to beneficiaries via medical diagnoses and national drug codes (NDCs). With the exception of the recently updated ICD-10-CM[1] codes, diagnosis-to-risk factor mappings rarely require updates. However, drug databases are updated and released regularly (i.e., monthly) if licensed through a vendor such as First DataBank (FDB). Figure 1 below depicts active monthly NDCs since the beginning of 2012, while Figure 2 portrays the incremental number of NDCs added or retired by month. Both charts use FDB's drug database to determine which NDCs are active (previously added and not terminated), newly added, and terminated by month.

Figure 1 – Number of Active National Drug Codes by Month



[1] International Classification of Diseases tenth revision to the Clinical Modification (ICD-10-CM) codes, effective October 1, 2015

Figure 2 – Number of Added and Retired National Drug Codes by Month



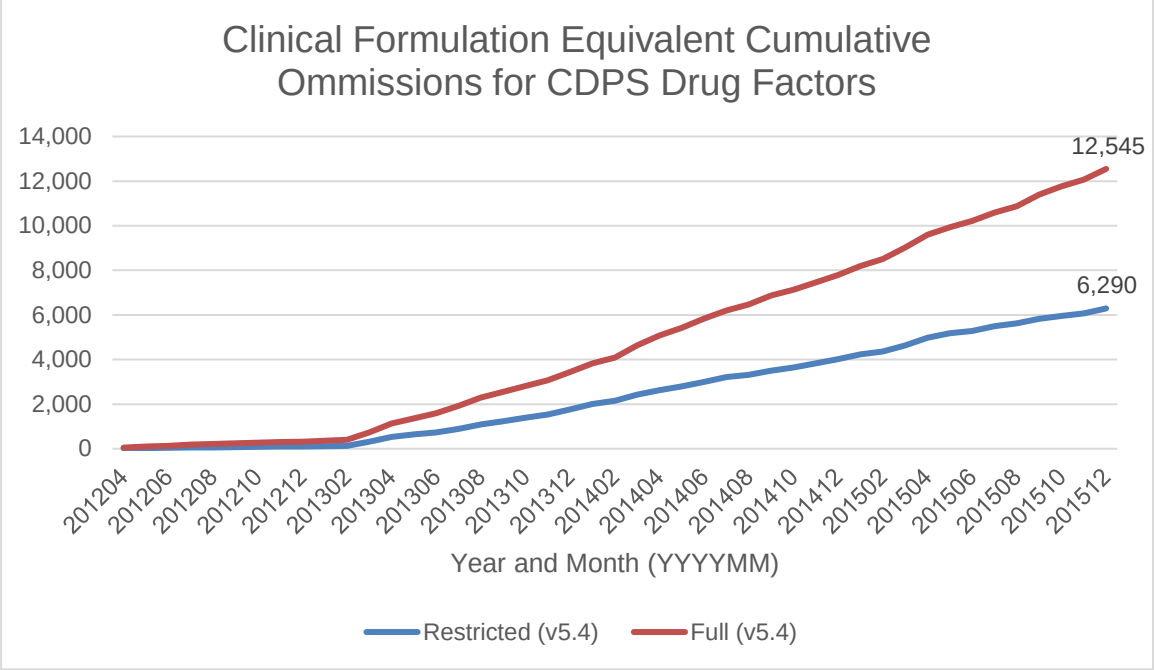
The most immediate observation is change: the number of NDCs changes each month either by addition or subtraction. Retired NDCs would hypothetically control for themselves in drug-to-CDPS factor mappings since prescribers would no longer use retired drugs. However, new drugs are concerning since prescribers would hypothetically move to more effective or lower cost drugs for which risk factor mappings do not receive regular updates.

Equivalent Drug Exclusions

FDB's MedKnowledge includes every drug approved by the FDA with description, pricing, and clinical decision support information. Clinical formulation equivalent codes allow users to identify identical drugs based on unique combinations of active ingredient(s), route, dosage form, and strength. These code sets have multiple uses including development of candidate drugs for substitution in the dispensing environment (FDB, 2016). Thus, we can assume that identical clinical formulations linking to existing NDC-based risk factors were unintentionally excluded or released after the most recent risk factor mappings.

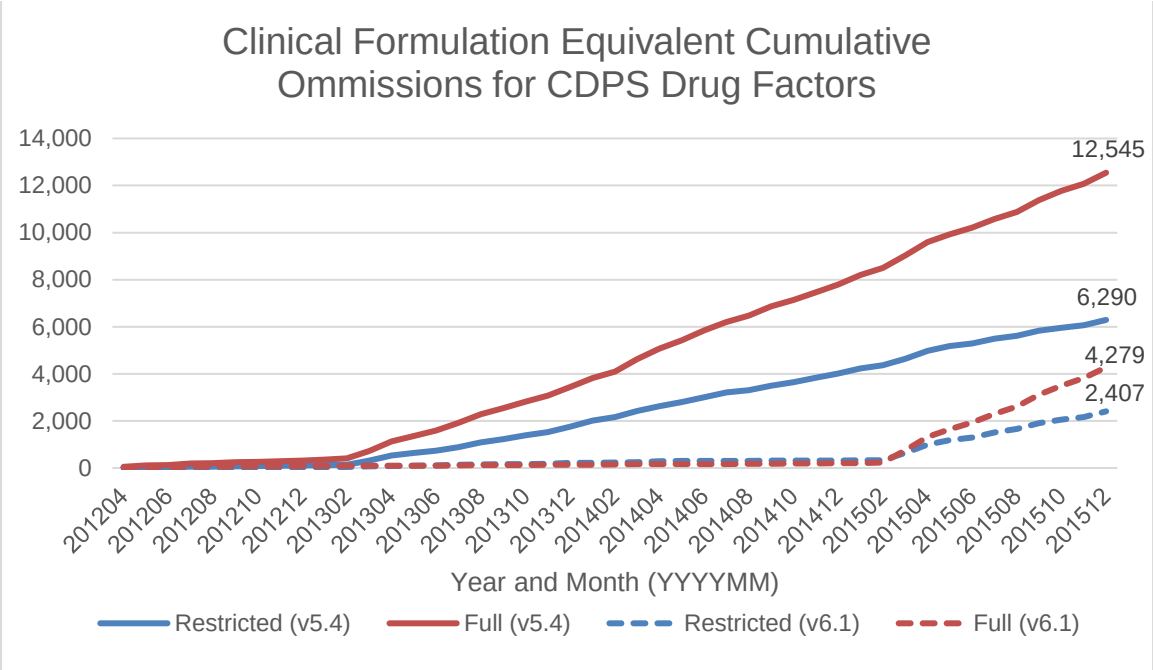
As depicted in Figure 3 below, a simple programmatic join (i.e., matching on clinical formulation code) of current NDCs *included within the algorithm* to generic equivalents *excluded from the algorithm* demonstrates the dramatic increase in missing equivalent NDC to CDPS factor mappings since the last algorithm update in early 2013. The Medicaid-Rx (full) model used in four states excludes more than 12,500 NDC codes (13% of current mappings) and rising while the combined CDPS (restricted) model using a limited number of pharmacy risk factors excludes more than 6,000 mappings (18% of current mappings).

Figure 3 – Number of Equivalent NDCs Excluded from CDPS Version 5.4 Risk Factor Mappings



Note that prior to the conclusion of this study, CDPS published version 6.1 of the software which includes updated pharmacy mappings. However, most states such as Texas continue to use previous software versions (TX currently uses version 5.4) due to rate setting deadlines and software release schedules. Figure 4 below reflects the impact before and after the version upgrade and simultaneously validates our approach to drug equivalent identification, since excluded NDCs were nearly eliminated from the observation period prior to March 2015. This would lead us to conclude the drug database used to derive version 6.1 drug mappings was last updated in February 2015 since the cumulative exclusion trend resumes directly thereafter.

Figure 4 – Number of Equivalent NDCs Excluded Before (v5.4) and After (v6.1) CDPS NDC Update



Based on current line slopes, we expect CDPS versions prior to 6.1 will exclude approximately a third more NDC mappings while version 6.1 excluded mappings will more than double by the end of 2016. We recommend two remedies for unintended risk factor omissions due to drug exclusions. First, we recommend an immediate upgrade to version 6.1 for programs using CDPS version 6.0 and prior. Second, for all risk programs using pharmacy data in risk score determination, including those using CDPS, we recommend an annual generic medication equivalence review and update prior to each model recalibration since it is unlikely that vendors are able to publish monthly updates to meet the varying schedules of state or national program periods. Drug equivalence validation is a relatively simple exercise, which should be inexpensive to perform either internally or externally.

Most Prescribed Excluded Drug Equivalents

Risk factor exclusions due to out-of-date NDC mappings are highly concentrated within very few drug-based risk factors. For the combined diagnostic and pharmacy model, two factors comprised more than 80% of excluded factors. The pharmacy-only model was expectedly less concentrated, yet the top two missing risk factors still comprised more than 50% of the total.

Table 2 – Most Common Equivalent Drug Exclusions

CDPS Model	CDPS Factor	CDPS Factor Description	% of Total Risk Factor Omissions*	Most Commonly Prescribed Drug(s) from Exclusion List
Restricted	MRX2	Cardiac	52.0%	Clonidine HCL
	MRX14	Seizure Disorders	32.1%	Lamotrigine, Gabapentin
Full	MRX25	Infections, Medium	33.8%	Cefinir, Clindamycin
	MRX37	Pain	20.1%	Multiple Combinations of Tramadol, Hydrocodone, and Acetaminophen

*Member Month Weighted

Note that excluded drugs passed additional validation efforts via inspection of FDB's Generic Therapeutic Classification (GTC) groupings for associated CDPS risk factor groups. CDPS "Cardiac" drug exclusions fell into two main groups, "Cardiac Drugs" and "Cardiovascular", while 100% of excluded "Seizure Disorders" drugs in the restricted model fell into the "CNS Drugs" (Central Nervous System) GTC. For the pharmacy-only (full) model, nearly 100% of excluded drugs fell into the "Antibiotics" and "Analgesics" GTCs for MRX25 and MRX37, respectively. Furthermore, within each GTC, prescribing behavior was typically highly concentrated on very few drugs. Lamotrigine, Gabapentin, and Lyrica comprised more than 91% of total utilization within excluded MRX14 drugs for the restricted CDPS model.

Risk Score Impact

Restricted CDPS Model

Risk scores for eligible enrollees were calculated before and after accounting for unintentional drug exclusions using Texas SFY15 calibrated weights. Paired t-tests are used in "before and after" studies to compare differences in two population means. These tests revealed statistically significant differences between actual and corrected risk scores for the total population, yet closer inspection reveals a seemingly trivial impact for enrollees less than 15 years old. Table 3 below depicts the population impact and risk score change of our member sample by program and risk group.

Table 3 – Restricted CDPS Model Impact by Program and Risk Group

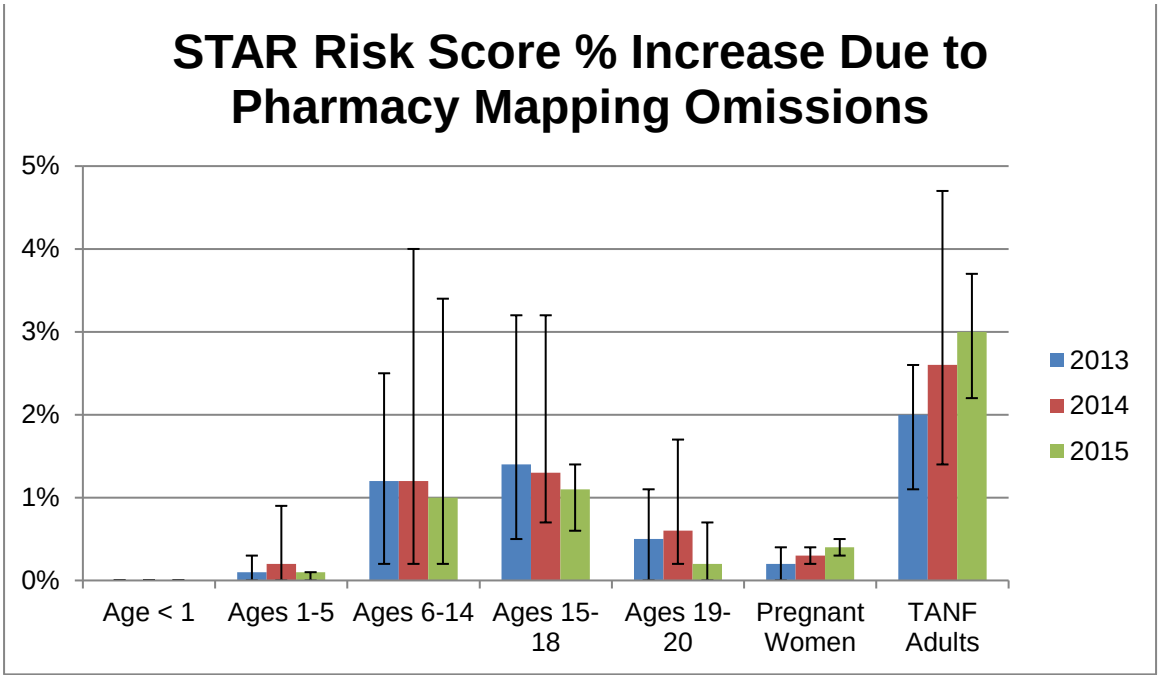
Program	Risk Group	% of Member Months Affected	% Risk Score Change
CHIP	Less than 1 Year of Age	0%	0%
	Age 1 to 5	0%	0%
	Age 6 to 14	0.2%	0.7%
	Age 15 to 18	0.8%	1.5%
STAR	Less than 1 Year of Age	0%	0%
	Age 1 to 5	0.1%	0.1%
	Age 6 to 14	0.3%	0.8%
	Age 15 to 18	1.0%	1.1%
	Age 19 to 20	1.0%	0.4%
	Pregnant Women	1.4%	0.3%
	TANF Adults	7.2%	2.6%

Recall that case mix ratios are calculated by program, risk group, service area, and plan. Therefore, calculating risk score changes at risk group levels does not truly illustrate the impact on rate setting since competing plans will utilize different drug benefits and prescribing patterns. While our enrollment sample does not include risk groups in overlapping service areas, we can still consider variance in risk score deficiencies at the program, risk group, and plan level.

The following two charts depict mean risk score changes (with the colored bars) across plans by fiscal year and risk group for each of the STAR and CHIP programs. Minimum and maximum plan risk score changes are also presented using error bars, which provide a more meaningful representation of adverse rate setting consequences due to excluded pharmacy mappings. Note that unlike the tabular results, risk score deficiency means are member month weighted at the individual plan level and not in aggregate. Minimum and maximum risk score changes in Figure 5 tell a different story than Table 3 above. For Ages 6-14, we see that minimum and maximum risk score changes vary significantly by fiscal year and signify that larger plans with lower risk score deficiencies are diluting higher risk score deficiencies from smaller plans in our sample. Recall from Table 1, this risk group represents the single largest program enrollment segment and such an increase in risk scores could have a very meaningful impact on premium revenue for this plan (not controlling for change in denominator of case mix formula).

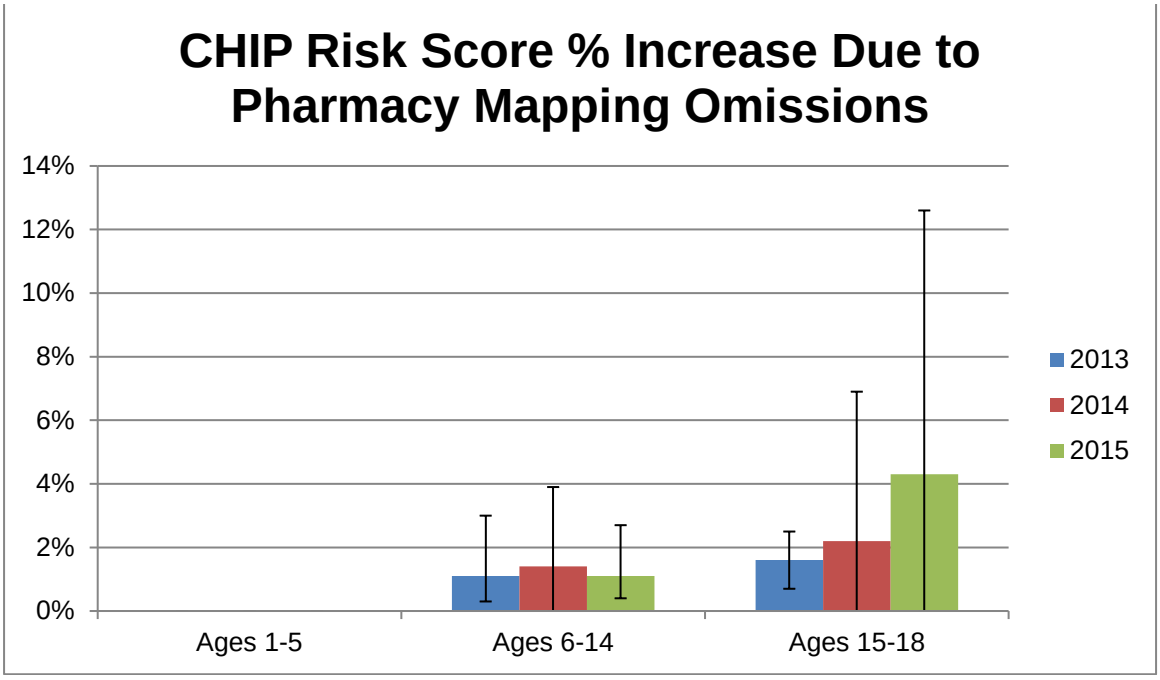
Spearman correlation coefficients can assist us in determining if risk score deflation is increasing with time. Correlation coefficients exist on a scale from -1 to +1, where +/-1 indicates a perfect (positive/negative) linear relationship and 0 indicates no relationship. These values were reviewed for risk score change and plan year by risk group. In this case, only Pregnant Women and TANF Adults produced a moderate linear relationship at nearly 0.5 for each risk group. Thus, we can conclude that risk score deflation is worsening over time (also depicted in Figure 5), yet only affects a subset of enrollees so far.

Figure 5 – STAR Program Risk Score Increase by Risk Group for Restricted Model



CHIP enrollees demonstrate greater variance than STAR program members, which is not an unexpected result with shrinking enrollment. A depiction of minimum, mean, and maximum plan level risk score deflation by risk group and year are presented for CHIP enrollees in Figure 6. Note changes in scale and omission of risk group "Less Than 1" due to insufficient enrollment across sampled plans.

Figure 6 – CHIP Program Risk Score Increase by Risk Group for Restricted Model



Full CDPS Model

The state of Texas uses only the combined diagnostic and pharmacy (restricted) version of the CDPS risk algorithm. However, we also assessed the impact of the pharmacy-only model for state Medicaid programs using the alternative model. Recall, we do not recalibrate and employ the risk factor weights as provided in version 5.4 of the CDPS algorithm. Note that some states may assess risk based on prospective risk scores, but our study limits results to concurrent weights only. Table 4 below quantifies the overall impact of drug mapping omissions by program and risk group. Affected member months and risk score change percentages expectedly raised sharply compared to the restricted CDPS algorithm since the full model is entirely dependent on pharmacy data to identify CDPS risk factors. The Pregnant Women category, previously affected less than other groups, now presents as one of the most affected groups when using only pharmacy data to quantify risk since adults take more medications than children. It follows that a relationship between age of the beneficiary and mapping omissions exists. Paired t-tests confirmed statistically significant differences between actual and corrected risk scores leading us to conclude current risk scores are calculated inaccurately.

Table 4 – Full CDPS Model Impact by Program and Risk Group

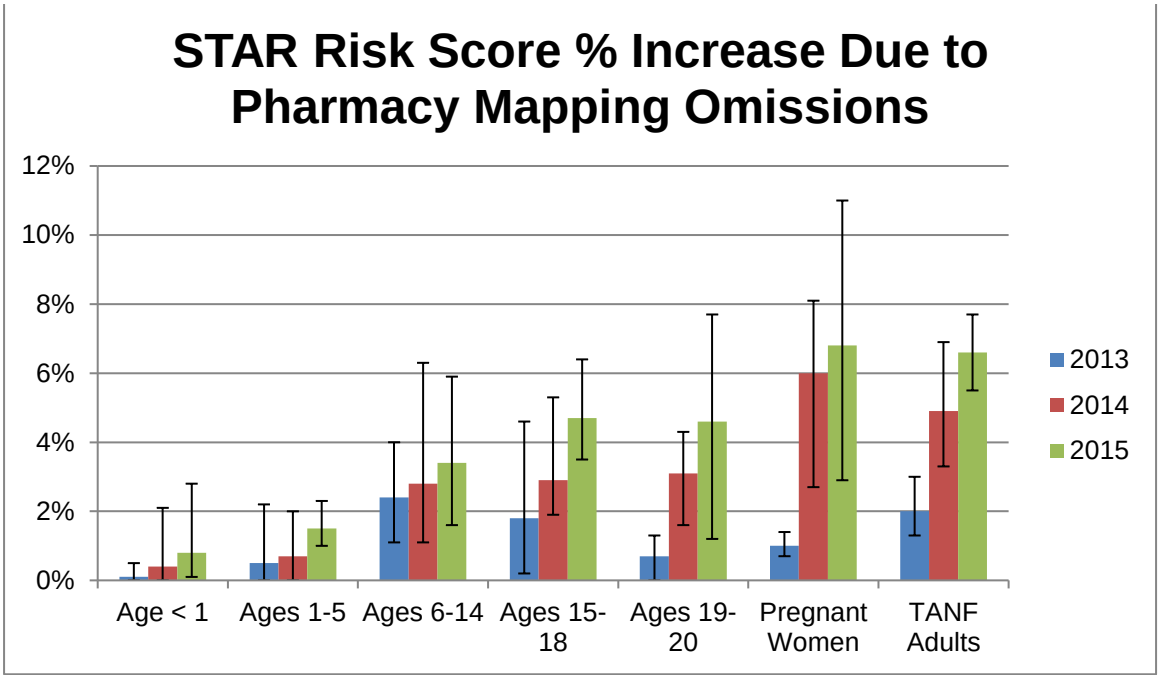
Program	Risk Group	% of Member Months Affected	% Risk Score Change
CHIP	Less than 1 Year of Age	0.2%	0.1%
	Age 1 to 5	0.8%	0.5%

Program	Risk Group	% of Member Months Affected	% Risk Score Change
	Age 6 to 14	0.8%	1.6%
	Age 15 to 18	1.5%	2.1%
STAR	Less than 1 Year of Age	1.2%	0.4%
	Age 1 to 5	1.2%	0.6%
	Age 6 to 14	1.6%	2.2%
	Age 15 to 18	2.9%	2.7%
	Age 19 to 20	5.4%	2.6%
	Pregnant Women	10.5%	3.6%
	TANF Adults	10.8%	4.5%

Furthermore, the issue of risk score deflation is increasing with each state fiscal year. While previously only detectible in TANF Adults and Pregnant Women, visual inspection of Figure 7 indicates a relationship of risk score deficiency and time across all risk groups. Spearman correlation coefficients were again reviewed to confirm this observation. Risk groups for enrollees less than 14 years old demonstrated weak linear relationships (0.3-0.5) and ages 15-18 demonstrated moderate linear relationships (0.5-0.7), while the remaining risk groups demonstrated very strong linear relationships with time (> 0.7).

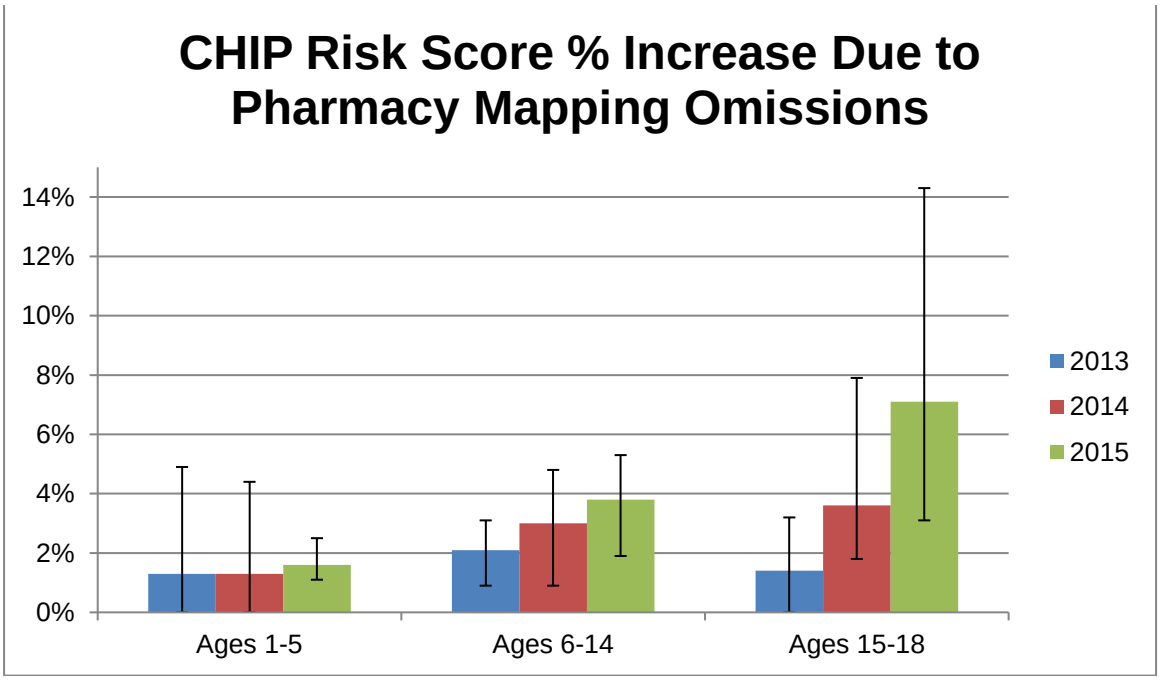
Plan level variance in risk score deficiency increased dramatically. Across all combinations of plan, risk group, and fiscal year for STAR enrollees, the largest corrected risk score increased by less than 5% under the restricted model while almost nine combinations exceeded that value under the full model with a maximum greater than 11%. We suspect such plan level variance would continue to exist across plans within identical service areas yielding over and underpayments via improperly calculated case mix ratios.

Figure 7 – STAR Program Risk Score Increase by Risk Group for Full Model



Similar to our previous illustration in Figure 6, CHIP enrollees demonstrate more variance than STAR program members for the full CDPS model with exception to enrollees aged 6-14.

Figure 8 – CHIP Program Risk Score Increase by Risk Group for Full Model



Recalibration

Model recalibration represents an important step to ensure model outputs reflect the claims input of the underlying population by modifying the risk factor weights. Annual recalibration is typical for most state and national plans incorporating risk adjustment in their rate setting process. Texas Medicaid employs a full recalibration approach by calibrating risk factor weights to the response variable without regard to the original model weights. While it is an essential duty of risk-based programs such as many state Medicaid organizations, recalibration has negative consequences if improperly applied.

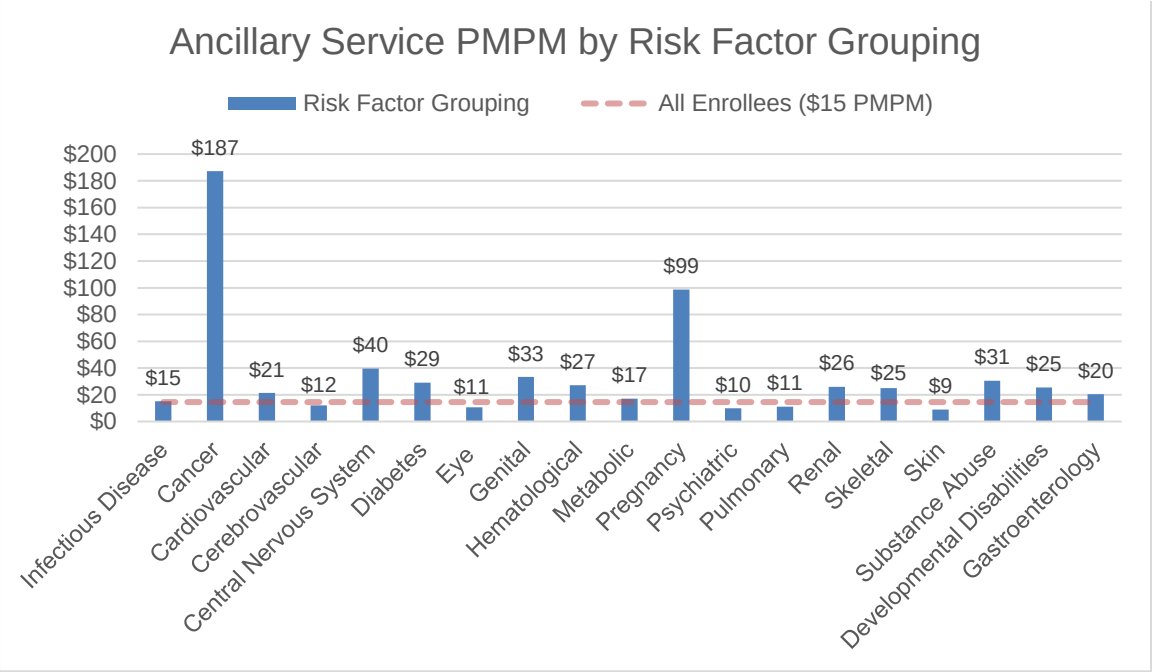
Dependent Variable

It is standard practice in risk adjustment application to exclude certain services from the determination of risk scores; most models require a face-to-face physician encounter to receive credit for risk factors. In Texas Medicaid, diagnoses linked to radiology and pathology procedures are omitted from calculation of risk scores. Similarly, the HHS-HCC risk adjustment methodology discusses the need to restrict diagnosis usage to claims where the procedure indicated a face-to-face visit with a qualified clinician. Services such as durable medical equipment, pathology, and radiology are not included (HHS, 2012).

The dependent variable in the HHS-HCC model is based on relative annualized plan liability expenditures using gross inpatient, outpatient, and prescription drug files (HHS, 2012). However, according to SFY 2015 Texas Rate Analysis (CDPS+Rx) Technical Specifications and confirmed by the EQRO, relative expenditures are determined absent of ancillary services in CPT ranges 70000-79999 (radiology) and 80000-89999 (pathology). While exclusion of ancillary services in risk factor attribution is standard practice, excluding ancillary services in relative expenditure calculation and recalibration is not required since it implies all risk factors carry proportional expenditures for ancillary benefits. Per conversation with the original author of CDPS, Todd P. Gilmer, Ph.D., Professor and Chief, Division of Health Policy for University of California, San Diego, he was not aware of such practice in recalibration and did not recommend exclusion of such services. His models excluded ancillary benefits from risk factor attribution, yet included the same services when modeling relative expenditures.

When analyzing ancillary spend types identified in the previous section, our enrollment sample required $15 total PMPM (allowed). However, calculating ancillary service PMPM by CDPS risk factor grouping illuminates the disproportionate spend in expenditures by CDPS risk factor. Figure 9 below groups CDPS factors into hierarchies designated within the CDPS algorithm and expresses ancillary services as a PMPM amount. We control for undeterminable dollar allocation across multiple risk hierarchies by selecting enrollees only within the specified hierarchy and no others. Thus, our results should reasonably quantify the typical PMPM expenditures expected for these combinations of risk factor groups and ancillary services. The chart immediately reveals individual risk groupings that far exceed the population average reference line. For example, cancer and pregnancy require an extremely high relative ancillary service PMPM and at the very least imply its risk factors are potentially undervalued based on our understanding of the recalibration methodology.

Figure 9 – Ancillary Service PMPM by Risk Factor Grouping



The issue is further compounded when MCOs consume disproportionate share of under- or overvalued risk groupings. Table 6 below presents health plan prevalence rates per 100,000 member months across programs (CHIP and STAR). In this instance, we remove the restriction requiring enrollees to have no comorbid risk groupings to illustrate the potential overall impact on plans. Note that a limitation of our study is that none of the surveyed plans overlap service areas. Thus, prevalence rates are presented for informational purposes only to demonstrate potentially adverse impact due to risk share variance. Table 6 illustrates a mitigating effect of our preceding cancer underweighting concerns due to extremely low prevalence rates. However, pregnancy and many other factors deviating significantly from the $15 PMPM expected ancillary costs demonstrate significant share variance, which implies plans are potentially incorrectly compensated based on the underlying risk profile. For pregnancy, we see that "Health Plan D" maintains nearly two-thirds less pregnant enrollees when compared to other plans. Should such a disparity exist within individual service areas, plans with disproportionate share of such populations would be undercompensated due to the much higher than average ancillary resource consumption.

Table 6 – Prevalence Rates per 100,000 Member Months by Risk Factor Grouping and Plan

Risk Factor Grouping	Health Plan A	Health Plan B	Health Plan C	Health Plan D
Infectious Disease	2,060	1,987	2,125	1,067
Cancer	179	168	270	241

Risk Factor Grouping	Health Plan A	Health Plan B	Health Plan C	Health Plan D
Cardiovascular	6,678	4,014	6,321	3,498
Cerebrovascular	223	178	95	242
Central Nervous System	2,561	2,097	3,545	1,841
Diabetes	1,124	853	1,176	580
Eye	471	411	551	388
Genital	1,330	1,410	1,432	853
Hematological	1,603	736	800	650
Metabolic	3,068	5,250	3,806	2,740
Pregnancy	4,640	4,504	5,204	1,650
Psychiatric	14,028	7,165	16,385	9,431
Pulmonary	21,407	13,249	16,736	15,963
Renal	1,487	1,233	1,428	921
Skeletal	7,652	7,477	7,496	6,574
Skin	7,140	4,698	6,321	4,929
Substance Abuse	448	450	575	197
Developmental Disabilities	130	232	166	194
Gastroenterology	5,419	4,783	5,812	3,766

Risk models commonly require recalibration to reflect its intended population's resource demands and deriving the dependent variable appropriately is the critical first step. Methodology misinterpretations such as extending the exclusion of ancillary services beyond the identification of enrollee risk factors represent recalibration hazards directly affecting risk-bearing entities financially. We recommend subsequent recalibration of all risk models require additional attention for dependent variable derivation to ensure such unintended consequences are avoided.

Independent Variables

Controlling for Insignificant Terms

Recalibrating models to new populations does not require previously accepted terms to remain statistically significant in new models. Geographic variation in disease prevalence and treatment protocols are two potential factors that can change the influence of risk factors. Additionally, recalibrating models originally devised from larger and perhaps more credible datasets using smaller datasets can invalidate the weights for rarer, more expensive conditions in the recalibration population. While these effects are not completely avoidable, there are steps we can take to mitigate the effects of unintended consequences due to small sample sizes or insignificant results.

First, let us understand the issue by surveying the recalibrated Texas risk factors for insignificant terms. Statistical significance thresholds vary in practice, but most statisticians use p-values less than either 0.05 or 0.01 to identify significant model terms. In our case, p-value measures the weight of evidence against term significance where a high value indicates strong evidence against the term (Wackerly et al., 2002). Thus, if we were to choose a modest acceptance threshold of 0.05, then terms with p-values greater than or equal to this value would indicate terms with insignificant value to the model. The following table represents recalibrated terms with p-values >= 0.05:

Table 7 – Insignificant Model Terms (Post Recalibration)

Texas Line of Business	Term	Description	Weight	p-value	Recalibration Sample with Condition
STAR Children	A_15_24F	15<=Age<25 Female	0.021	0.26	322,032
	MRX8	HIV (medication)	0.048	0.90	434
	MRX15	Tuberculosis (medication)	0.197	0.23	1,929
CHIP	A_5_14F	5<Age<15 Female	-0.003	0.86	535,637
	A_5_14M	5<Age<15 Male	0.014	0.44	560,736
	A_15_24F	15<=Age<25 Female	-0.024	0.27	211,432
	BABY3	Mild Prematurity (32-36 Weeks)	-0.153	0.71	317
	BABY7	Single, Term Infants Without Problems	1.364	0.08	95
	BABY8	Twin Infants	0.47	0.91	< 10
	MRX15	Tuberculosis (medication)	0.192	0.49	627

Aside from age and gender terms, six terms exist within two distinct risk factor sets without statistical significance. We can confidently state that terms approaching 1.00 have no significance to the resulting

model. HIV medications in STAR children is one such term with near certainty of insignificance, which creates additional problems in the interpretation of clinical hierarchies that will be further discussed in a subsequent section. Let us instead focus on two terms within the CHIP population that create issues when explaining risk adjustment to a non-technical audience: BABY3 and BABY7. Clinicians and health plan administrators alike often have difficulty understanding negative terms such as BABY3 since the literal interpretation is that a premature birth makes the enrollee less costly, not more. The issue compounds when paired with other terms in the model such as normal newborns (BABY7), which implies inducing delivery four to eight weeks prior to term will save the health system money. The issue is most likely explained by collinearity, i.e., mildly premature infants contain additional risk factors that increase overall risk scores, which are not immediately obvious through interpretation of model results.

While these terms carry low weights or affect small populations, it is important to maintain best practices during recalibration to ensure model results are explainable to broader audiences and are not biased to meager sample sizes. Plan expenditures in new geographies not part of the calibration sample could yield vastly different results should prevalence not mirror prior populations. We recommend two candidate strategies to handle insignificant terms:

1. Full Recalibration: full recalibration requires calibrating risk factor variables to the response variable without regard to the base model weights. In this instance, we suggest excluding insignificant terms and refitting the model with the remaining terms. If traditional measures of fit remain unchanged (or improved), then sufficient evidence exists to accept the new model absent of excluded terms.

2. P-Value Weighting: this approach was discussed in the 2007 SOA sponsored study, *A Comparative Analysis of Claims-Based Tools for Health Risk Assessment* (Winkelman and Mehmud), and more recently in an issue of *Predictive Analytics and Futurism Newsletter* (Hileman). Instead of regressing against the relative cost ratio, this approach regresses to the residuals of the predicted base model and actual relative cost ratios. The resulting p-values are used to determine credibility weights by subtracting the values from one. Thus, insignificant p-values (approaching one) will appropriately render less influence on recalibrated scores, which are obtained by summing the originally predicted scores with the dot products of the regression coefficients, the credibility weights, and the independent variables.

Preservation of Clinical Hierarchies

CDPS risk factors are defined using clinical hierarchies of severity intending to identify which risk factors require more resources to manage within each grouping of terms. These hierarchies are typically designated by terminology such as “Extra High”, “Very High”, “High”, “Medium”, “Low”, “Very Low”, and “Extra Low” within the risk factor descriptions. Clinical hierarchies with more than one term are defined below in order of hierarchy rank where the first term is of highest rank:

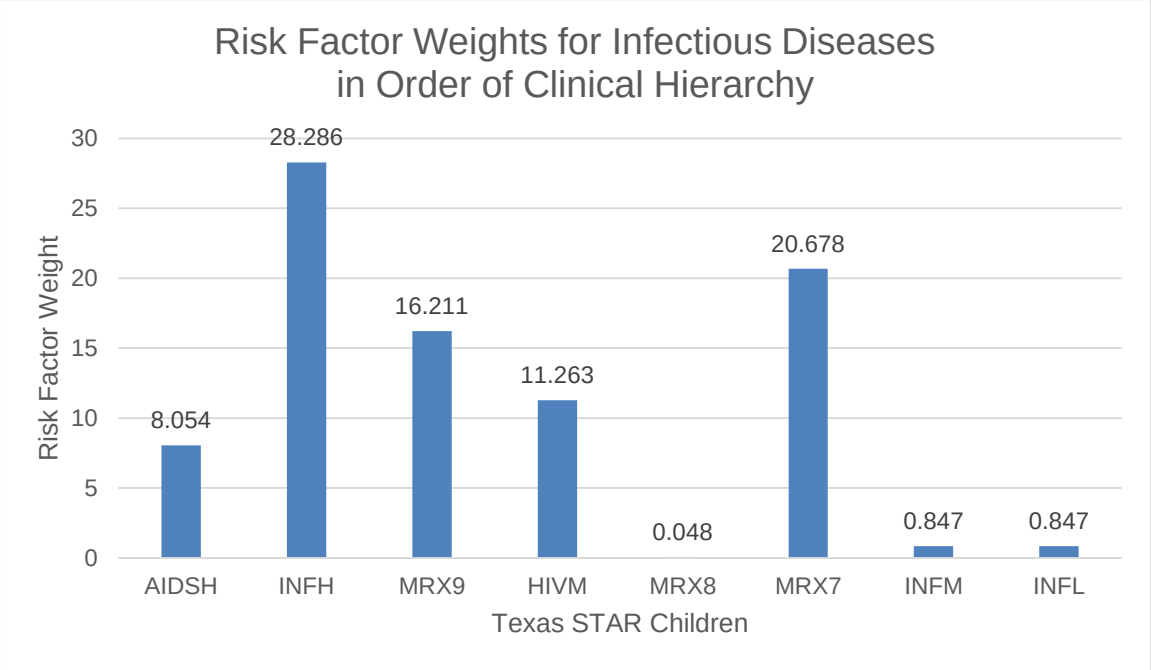
Table 8 – CDPS Risk Factor Hierarchies in Order of Rank

Risk Factor Grouping	Risk Factor Hierarchy Order
Infectious Disease	AIDSH, INFH, MRX9, HIVM, MRX8, MRX7, INFM, INFL
Cancer	CANVH, CANH, CANM, MRX11, CANL
Cardiovascular	CARVH, CARM, MRX1, CARL, CAREL, MRX2
Central Nervous System	CNSH, CNSM, MRX12, CNSL, MRX14, MRX13
Diabetes*	DIA1H, DIA1M, DIA2M, DIA2L
Eye*	EYEL, EYEVL
Hematological	HEMEH, MRX6, HEMVH, HEMM, HEML
Metabolic	METH, METM, METVL
Pregnancy	PRGCMP, PRGINC
Psychiatric	PSYH, PSYM, PSYML, PSYL, MRX3
Pulmonary*	PULVH, PULH, PULM, PULL, MRX15
Renal	RENEH, RENVH, MRX5, RENM, RENL
Skeletal	SKCM, SKCL, SKCVL, MRX10
Skin	SKNH, SKNL, SKNVL
Substance Abuse	SUBL, SUBVL
Developmental Disabilities*	DDM, DDL
Gastroenterology	GIH, GIM, GIL

*These groupings potentially collapse hierarchical terms into lower severity terms (dependent on enrollee attributes). The remaining terms maintain a clinical hierarchy with fewer terms.

As a consequence of recalibration, clinical hierarchies are potentially broken and result in lower ranked terms receiving larger weights than terms with superior clinical significance. Infectious diseases is one such example of hierarchy voidance and one of the more complicated hierarchies to manage since it contains the most individual risk factors (eight). Figure 10 presents the recalibrated weights for Texas STAR children with infectious diseases. A preservation of clinical hierarchies would be represented by descending risk factor weights from left to right. Thus, a term with a risk factor weight greater than any term to its left represents a voidance of clinical hierarchy.

Figure 10 – Texas STAR Children Weights for Infectious Diseases

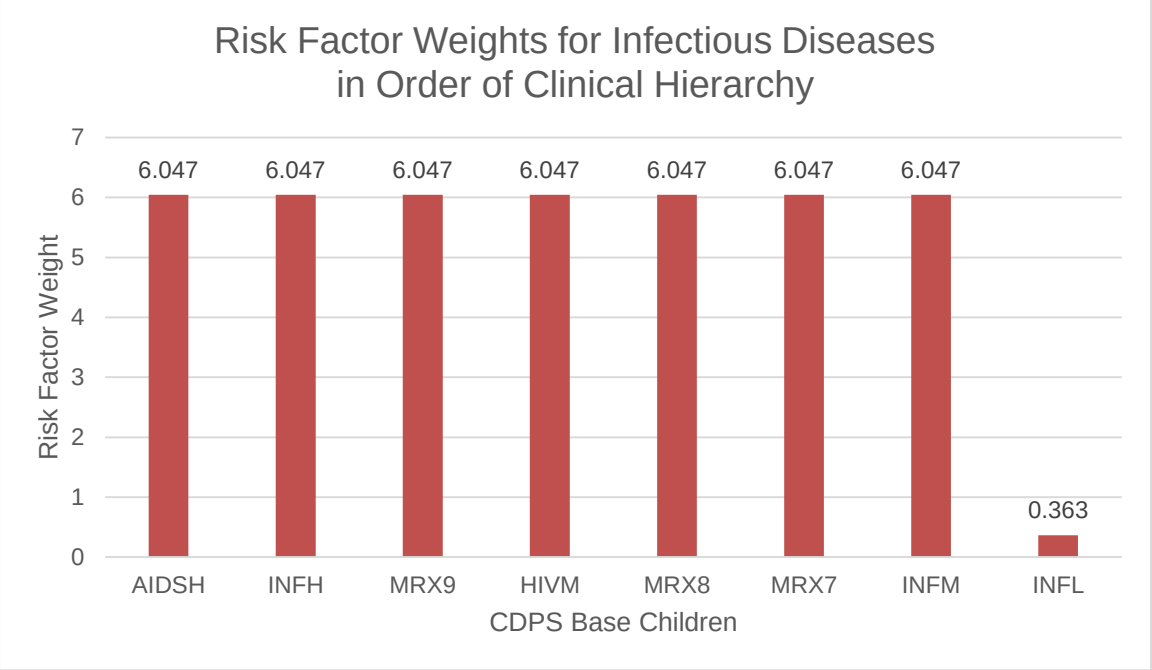


Several interpretations follow immediately from the chart above; namely, four CDPS model risk factors (INFH, MRX9 HIVM, MRX7) are suggested to be costlier than the highest ranked term, AIDS, High. In a review of Texas recalibrated weights, we found seven, six, and seven broken hierarchies for STAR Adult, STAR Child, and CHIP populations, respectively.

We also discussed these concerns with Dr. Gilmer (CDPS developer) who indicated best practice in recalibration necessitates a SAS ***RESTRICT*** statement applied within the regression statement to preserve the clinical hierarchies. Otherwise, collinear terms, biased samples, or outliers can potentially void clinical hierarchies and imply that clinically less intensive conditions require more resources than conditions with higher severities. Note, we also recommend use of the same statement to ensure risk scores are non-negative; otherwise while extremely unlikely, the possibility of a negative plan risk score exists, which would imply a plan's program, risk group, and service area combination would pay premium for each member month of enrollment instead of receiving. Parameter estimates can be corrected by adding restrict statements in SAS regression procedures in the form of ***restrict equation1=equation2*** (SAS/STAT User's Guide, Version 8, 1999). In the above example, we would initially start with a statement in the form of restrict MRX8=MRX7. If the statement resulted in terms for MRX8 and MRX7 that continued to exceed that of HIVM (or higher ranked terms), we would continue adding higher ranked terms and rerunning the

procedure until we were left with no terms violating the hierarchy. Evidence of such practice exists in the base calibrated weights provided by Dr. Gilmer at the time of the interview in Figure 11:

Figure 11 – CDPS Base Children Weights for Infectious Diseases



The chart above demonstrates best practice in recalibration and we recommend an enforcement of clinical hierarchy preservation using the steps previously described in this section.

Conclusions

Our case study of risk adjustment for Texas Medicaid demonstrates the need for regular updates to pharmacy based risk factor mappings using NDC codes. If possible, these risk factors should be updated during each rate setting period to ensure that MCO risk scores accurately reflect the disease burden of its beneficiaries. Otherwise, risk scores missing formulary equivalent drug mappings will be deficient when prescribers move to newer drugs. The issue further compounds with time; due to the frequent release of new drugs, longer periods between updates will exhibit greater risk score deficiencies.

External review organizations and actuaries should also exercise caution during recalibration. Risk models carry assumptions about the input data such as the intention to predict total spend. When recalibrating to alternative assumptions, risk factors may exhibit disproportionate weights to the risk adjustment specifications outlined during rate setting. Finally, maintaining clinical hierarchies and proper model acceptance criteria is critical to sustaining credibility amongst the MCOs ultimately measured by risk models. Model coefficients that change dramatically over risk adjustment periods and less severe conditions carrying higher weights than clinically recognized more severe conditions can create confusion amongst risk adjustment adopters, which can be avoided through adoption of best practices.

Glossary of Terms

CDPS: Chronic Illness & Disability Payment System
CHIP: Children's Health Insurance Program
Dx: Diagnosis
EQRO: External Quality Review Organization
FDB: First DataBank
GTC: Generic Therapeutic Classification
HCC: Hierarchical Condition Categories
HHS: Health and Human Services
ICD: International Classification of Diseases
MCO: Managed Care Organization
NDC: National Drug Code
PMPM: Per Member Per Month
Rx: Pharmacy
SAS: Statistical Analysis System (software)
SFY: State Fiscal Year
SOA: Society of Actuaries
TANF: Temporary Assistance for Needy Families
TX: Texas
UCSD: University of California, San Diego

References

Express Scripts, Issues Document: Patent Expirations (2013-2017), Express Scripts, 2013. Retrieved from http://www.centerlighthealthcare.org/images/uploads/Brand_Name_Drugs_with_Patent_Expirations_2013_-_2017.pdf.

World Health Organization. *International Classification of Diseases (ICD)*, World Health Organization, 2016. Retrieved from http://www.who.int/classifications/icd/en/.

First Databank, Inc (FDB). *FDB MedKnowledge*, First Databank, Inc., South San Francisco, CA. Accessed May 2016.

Department of Health and Human Services. *Provisions of the Proposed HHS Notice of Benefit and Payment Parameters for 2014*, Federal Register, Vol. 77, No. 236, Part II. U.S. Government Publishing Office, 2012.

The Institute for Child Health Policy. *Technical Specifications TEXAS Rate Analysis (CDPS+Rx)*, State Fiscal Year 2015, University of Florida, March 29, 2016.

Gilmer, Todd P. *Chronic Illness & Disability Payment System*. University of California, San Diego. Web. 24 May 2016. Retrieved from http://cdps.ucsd.edu.

SAS/STAT User's Guide, Version 8. Cary, NC. SAS Institute Inc., 1999. Print.

Wackerly, Dennis D., William Mendenhall, III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. 6th ed. Pacific Grove, CA. Duxbury, 2002. Print.

Hileman, G. A Comparison of Risk Scoring Recalibration Methods. *Predictive Analytics and Futurism Section Newsletter*, December 2015. Retrieved from https://www.soa.org/Library/Newsletters/Predictive-Analytics-and-Futurism/2015/december/paf-iss12.pdf.

Winkelman, R., and Mehmud, S. *A Comparative Analysis of Claims-Based Tools for Health Risk Assessment*. Society of Actuaries, 2007. Retrieved from https://www.soa.org/research-projects/health/hlth-risk-assessment.aspx.