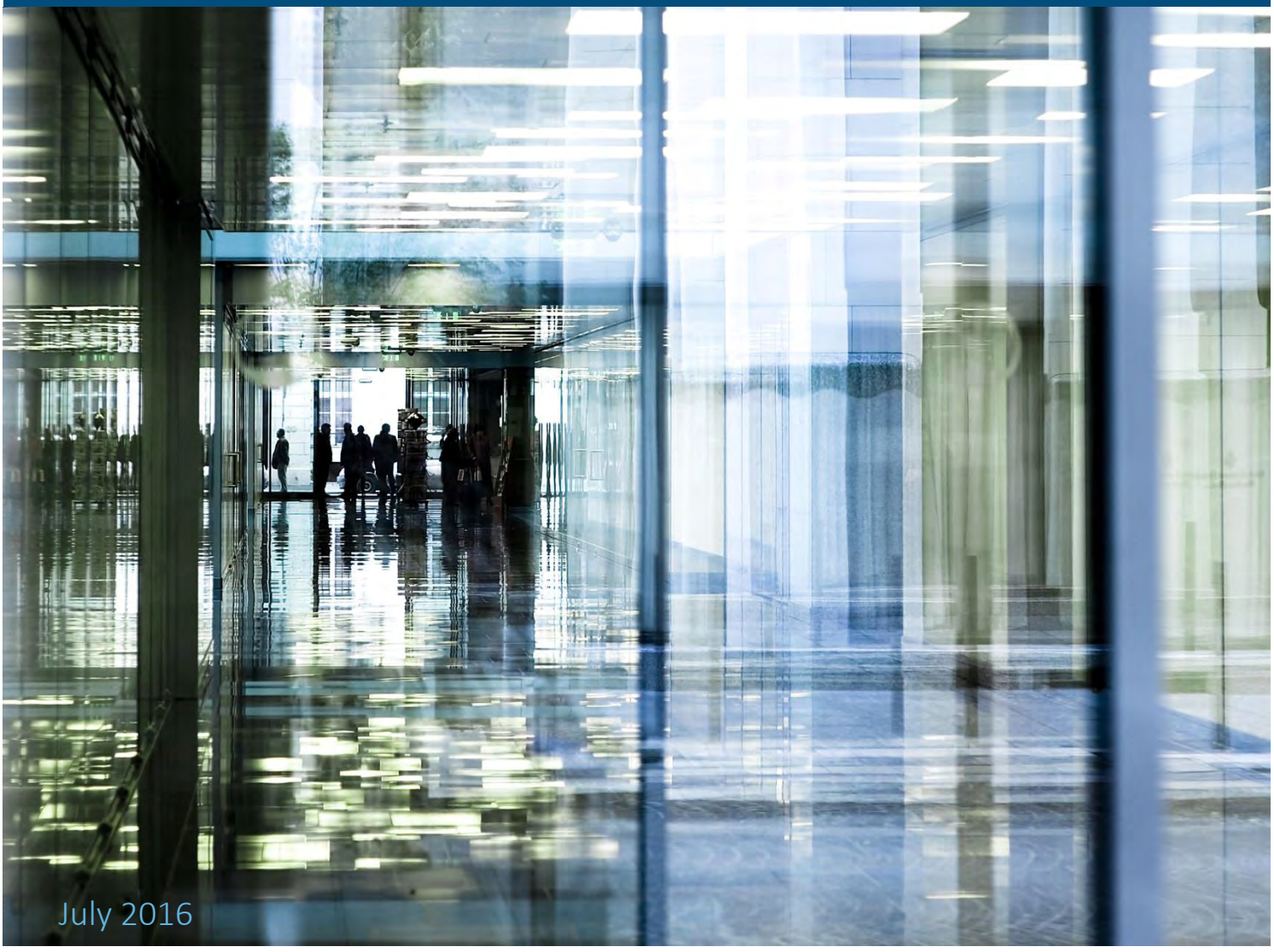
SOCIETY OF ACTUARIES

Risk Scoring in Health Insurance:
A Primer



July 2016

Risk Scoring in Health Insurance:

A Primer

SPONSOR Society of Actuaries Health Section

AUTHORS

Geoffrey R. Hileman, FSA, MAAA
Kennell and Associates, Inc.

Syed Muzayan Mehmud, ASA, FCA, MAAA
Wakely Consulting Group, Inc.

Marjorie A. Rosenberg, PhD, FSA
University of Wisconsin–Madison

Caveat and Disclaimer

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information.

TABLE OF CONTENTS

Acknowledgments

We are grateful to the Society of Actuaries and the Health Section Research Committee in particular for their support in funding this research effort. We are also very appreciative of Steven Siegel and Barbara Scott for their administration of the project. We thank the Project Oversight Group, who provided valuable feedback throughout the development of this report. The Group comprised John Bertko, Kristi Bohn, Cabe Chadick, Ian Duncan, Rafi Herzfeld, David Knutson, Rick Lassow, Tom Messer, Rebecca Owen, and Bernie Rabinowitz. Finally, we also wish to recognize Veronika Badurova, Dave Kennell, and Spenser Steele of Kennell and Associates for review and technical assistance at various stages of the project.

Section 1: Introduction

Many of the extensive market reforms resulting from the 2010 Affordable Care Act (ACA) became effective beginning in calendar year 2014. These reforms were designed to increase the number of individuals covered by health insurance, create insurance coverage that was both comprehensive and affordable, and ensure minimum levels of coverage (through mandated benefits) and easily comparable premiums (through the establishment of the "metal" structure). This was intended to facilitate increased consumer understanding of the marketplace and to provide consumers with a framework more tailored to comparing across plans. Three key elements were included in the ACA to achieve these goals from the consumer's perspective: (1) an individual mandate that introduced a financial penalty for failing to purchase health insurance; (2) premium subsidies and reduced out-of-pocket costs for lower income families; and (3) rating reforms to ensure that premiums were more consistent among individuals, including unisex rating, the compression of the age curve for premium determination, and the elimination of both medical underwriting and preexisting conditions restrictions.

As insurers are no longer permitted to medically underwrite (that is, to decline coverage or to differentiate rates based on medical conditions) individuals seeking health insurance under the ACA, additional reforms were necessary from the perspective of the insurer to mitigate the substantial uncertainty of these newly insured risks. New mechanisms to properly price and manage the financial impact of policies issued were created and implemented as a result of the ACA. The three most visible of these reforms are commonly referred to as the "3 R's": reinsurance, risk corridors and risk adjustment. The purpose of this paper is to examine the fundamentals of risk scoring and its implementation in the post-ACA commercial marketplace and other applications, including Medicare and Medicaid.

The risk adjustment program is the third component of the premium stabilization programs for the ACA and, unlike reinsurance and risk corridors, is a permanent program. The goal of the risk adjustment program is to adjust payments to insurers to reflect the actual risk profile of the individuals who enroll in their plans relative to other plans in the same state and block. The risk adjustment program is divided into two stages. The first stage is the determination of a "risk score" of each insured population. The second stage is the risk transfer formula that is used to balance the premiums among the health plans to reflect differences in risk scores of the enrolled population by health plan.

The purpose of this paper is to provide a more detailed exploration of the first stage of the risk adjustment program, the risk score model that we refer to as "risk scoring." We discuss the history and considerations related to risk scoring beyond its application in the ACA context. For readers interested in exploring the risk transfer formula in more detail, see Pope et al. (2014) as one reference.

Although today's actuaries may understand that the term "risk adjustment" relates to a system or program to transfer payments among providers or insurers, the use of this term is not the same across all disciplines, nor over time. Iezonni (2012) notes that historically clinicians used the terms "severity" and "risk" as synonyms. She observes that when the DRGs were introduced in the 1980s that "risk adjustment" became intertwined with words such as "case mix," "severity," "sickness," "intensity," "complexity," "comorbidity" and "health status" by clinicians, policymakers, payers and actuaries. Even the first Society of Actuaries sponsored monograph "A Comparative Analysis of Methods of Health Risk Assessment" referred to "risk assessment" as the first stage of the process to estimate the risk, and referred to "risk adjustment" in the second stage as the payment transfer mechanism (Dunn et al. 1996).

In this paper, we focus on the "risk scoring," which we define as the first stage of adjusting for the differing level of risk. The combination of risk scoring plus the second stage process of any transfer of payments is referred to as "risk adjustment." In Section 2 we provide an overview of risk scoring, along with its history, and how risk scoring can result in mitigating uncertainty. We indicate that risk adjustment under the ACA is not a new concept, but one that modifies other risk adjustment programs that were or are in use. We also discuss other possible approaches to risk scoring and examine why the current method was chosen. In Section 3 we explore other areas where risk scoring has been used, such as in Medicare and Medicaid programs.

Section 2: Overview of Risk Scoring

Risk scoring, or adjusting for the severity or case mix of a population, has been studied for more than 40 years as a way to provide meaningful results in studies involving health care data (Iezzoni 1997b). The end use of any risk scoring study is to measure outcomes whether at the patient level, provider level, hospital level or population level in studies of mortality rates, utilization or expenditures in total or for a particular disease, in creating provider report cards, or in measuring various aspects of quality. As Iezzoni stated with regard to hospital inpatient death rates in 1997 (using the term *risk adjustment* for our risk scoring):

> The rationale for risk adjustment is to remove one source of variation, leaving residual differences to reflect quality. The underlying assumption is that outcomes result from a complex mix of factors: patient outcomes equal effectiveness of treatments plus patient risk factors that affect response to treatment plus quality of care plus random chance. (Iezzoni 1997b)

The specific choice of patient risk factors included in a risk scoring model may result in unintended consequences. For example, if certain characteristics, such as race or income, were included in the risk model, and if medical disparities were present, then the outcome result would not reflect these disparities (Romano 2000). Also, a risk scoring model designed for one outcome (like death) may not be a suitable risk scoring model for another outcome (like medical expenditures) (Iezzoni 1997b). Without the use of risk scoring, insurers would have a strong incentive to avoid high-cost members. However, these high-cost members may benefit from programs that health plans offer to improve their health (Kronick et al. 1998).

Results from a risk scoring model rely on a set of key risk scoring principles (Iezzoni 2012). These include (1) the purpose of conducting the study (such as studying death, utilization, costs, quality or efficiency), (2) population of interest, (3) time frame of study, (4) data source, (5) variables/factors included, (6) model development approach (whether statistical, clinical or both), (7) acceptable level of accuracy of results and (8) implementation approach. Each of these key principles is discussed throughout this paper.

2.1 History of Risk Scoring

The origins of risk scoring can be traced back to the 19th century with Florence Nightingale (1820–1910) and William Farr (1807–1883), a physician, who suggested that hospital mortality statistics of England's hospitals may have been misleading because differences across patients were not considered (Iezzoni 2012). The United States government, as part of the 1972 Social Security amendments, established the Professional Standards Review Organizations (PSROs) to monitor differences among medical providers. Through this legislation, the federal government established utilization review, statistical profiles and medical audits to maintain a balance between appropriate, effective and quality medical care (Donabedian et al. 1982).

Early risk scoring mechanisms were created to respond to specific requests, such as the Computerized Severity Index (CSI), MedisGroups or DiseaseStaging (Iezzoni 1997b). Tables 1.3–1.8 of Iezzoni (1997a) provide a summary of some risk scoring models and list the definition of severity, the pertinent patient population, the role of diagnosis, the role of major surgery, data requirements, how the measure was developed (whether clinical, empirical or both), timing of measurement and classification approach.

Many of the early risk scoring tools were created to compare hospital inpatient mortality rates. Prior to the existence of these risk scoring tools, the evaluation of differences across hospitals included the use of implicit, or unwritten, protocols depending solely on a reviewer's professional training, experience and judgment. Early research on the use of these implicit protocols showed them to be unreliable (Zimmer 1974; Rosenberg 1975). The advantages of explicit over implicit protocols are that they are "objective, verifiable, uniform across different hospitals, physician specialties, and types of patients" (Payne 1987). The use of these explicit protocols led to the development of more empirical or statistically based protocols.

Statistically based methods are based on the use of meaningful classification of diseases, with groupings differing depending on the purpose of the study. As discussed in Feinstein (1967) the underlying assumptions of a disease classification system include the following:

i. The medical care process is a reproducible, science-based act, related to the etiology or manifestation of the symptom. While the process is subject to large individual patient and physician variations, it is possible to construct protocols reflecting consensus of the appropriate behavior of the physician. This is the foundation of 'scientific medicine'.

ii. Diseases can be classified into reliable, nominal groups, that is, a patient exhibiting certain characteristics would be consistently placed into the same disease group by different coders.

iii. The elements of the nominal scale can be aggregated efficiently for various purposes, and the taxonomy will prove useful for those purposes.

The underlying disease classification system that is the foundation of US risk scoring models is an extension of the World Health Organization's (WHO) International Classification of Disease (ICD). Internationally, the first version of the ICD system was in 1900; previous approaches to classification have existed since 1785 (WHO 1975). The ICD classification system had its origins for the development of mortality statistics and not for use in morbidity classification systems. The ninth version of the ICD was a major revision to consider the classification of disease for both mortality and morbidity considerations: "The ICD-9 is designed for the classification of morbidity and mortality information for statistical purposes, and for the indexing of hospital records by disease and operations, for data storage and retrieval" (NCHS 1979).

The US National Center for Health Statistics (NCHS), the primary health statistical arm of the US Department of Health and Human Services, is responsible for supporting the ICD. The clinical modification of the ninth revision of the ICD (ICD-9-CM) "serves as a useful tool in the area of classification of morbidity data for indexing of medical records, medical care review, and ambulatory and other medical care programs, as well as for basic health statistics. To describe the clinical picture of the patient, the codes must be more precise than those needed only for statistical groupings and trend analysis" (NCHS 1979).

The ICD-9-CM disease classification system formed the basis of the risk scoring models used in programs throughout the United States. After years of preparations and delays, the successor to ICD-9-CM, ICD-10-CM, was implemented in the United States on October 1, 2015. ICD-10, under development since the early 1980s, represents a significant expansion in the level of detail that is available in diagnostic coding. One example of this increased specificity is in codes for pressure ulcers. Under ICD-9, there are nine codes indicating pressure ulcers that describe the broad location of the ulcer. The ICD-10 code set includes 150 different pressure ulcer codes, with more specific locations and the depth, or severity, of the ulcer. As coding practices eventually fully utilize this more specific code set, future generations of risk scoring models may be capable of more precise estimation of risk of individuals. In the short term, most risk scoring models will map ICD-10 to ICD-9 codes, and thus it is not likely that there will be an immediate increase in explanatory power.

2.2 Technical Concepts Related to Risk Scoring Models

In this section, the basics of a risk scoring model are discussed. We assume that the purpose of the study and the population of interest, as listed in the risk principles above, are defined. Here we lay out the general statistical framework (linear regression), consider the timing of the prediction, consider whether it is in the same year (concurrent) or whether the predicted risk score is in the next year (prospective), discuss in greater detail the choice of the explanatory variables included, and explain how to use the risk scoring model for other populations than what is originally designed (called "recalibration"). Some technical areas of how to handle individuals with incomplete data and the limitations of the risk scoring model conclude this section.

2.2.1 Introduction

Fundamentally, risk scoring is the process by which information about an individual is used to predict, or explain, uncertain events. These events may include mortality rates, utilization of health care services, health care efficiency or health care costs. The application of risk scoring to predicting and explaining health care expenditures is the primary emphasis of this paper.

Risk scoring models make use of a mathematical function to link individual attributes, such as demographics and diagnoses, and produce an estimate of annual health care expenditures. The type of expenditure represents either the total amount incurred by an individual regardless who is financially responsible to pay for the care (the allowed cost) or the amount that is the responsibility of the plan or payer (the paid cost).

For use in actuarial risk models, the expenditure variable is transformed to a risk score by rescaling the expenditure to a mean of 1.0. Let the pair of (z_i, y_i) be the expenditure and risk score for the ith individual respectively. As shown in Equation 1 the rescaled cost for individual i is found by dividing the expenditure of individual i by some mean cost. The mean cost could be calculated from the sample of individuals on which the model is being estimated, or the mean could be based on some exogenous reference population of individuals to which the model would be applied:

$$y_i = \frac{z_i}{\bar{z}} \quad (1)$$

Risk scores with this design are easy to interpret, as a score of 1.0 is equivalent to a person whose health care costs are exactly equal to the population mean, while a risk score of 0.1 would indicate that an individual has expected expenditures equal to 10% of the average.

Risk scoring models are frequently specified as linear regression models. Equation 2 illustrates the general framework for a regression-based perspective on risk scoring. Let y_i be the risk score for the ith individual, x be the vector of p explanatory variables, β be the vector of $p+1$ unknown regression parameters, and ϵ_i be the error term for the ith individual:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (2)$$

The independent variables in the risk scoring model are the patient-specific risk factors or individual attributes. Most commonly included explanatory variables include demographic factors, indications of the presence of specific medical conditions or groups of conditions, and indicators for the use of certain prescription drugs. For example, an entire sequence of diagnosis codes could be mapped into a single category indicating a broader condition, such as "cardiovascular disease." Some models may include data other than diagnoses as independent variables, potentially including medical procedures undergone by the individual, socioeconomic status, clinical lab or biometric data, or lifestyle data.

As in linear regression model estimation, the errors are assumed to be approximately normally distributed, and least squares methods are used to estimate β, the vector of unknown regression coefficients. The predicted risk score is calculated by multiplying the vector of fitted regression coefficients associated with each of the independent variables.

2.2.2 Timing of Prediction: Concurrent and Prospective Models

The dependent variable is the risk score as some function of health expenditures. Depending on the purpose and use of the model, the risk score is estimated for either the current period or a future period. These are commonly referred to as concurrent (estimating risk in the current period) or prospective (estimating risk in a future period) models. The period is typically 12 months but could be specified as some other defined time.

If the purpose of the model is primarily of an explanatory nature, then a concurrent model is used. A concurrent model uses the explanatory variables from the same base period as the risk score. The types of explanatory variables are discussed in detail in Section 2.2.3. As an example, a concurrent model to estimate calendar year 2015 risk scores uses diagnoses observed in calendar year 2015 as explanatory variables (the vector x from Equation 2).

If the purpose of the model is for prediction, then a prospective model is used. A prospective model defines the explanatory variables from a base period to predict the risk score in a later target period, typically the next one. For example, a prospective model might use diagnoses observed on medical encounters in calendar year 2014 to predict risk scores in calendar year 2015. While prospective models typically use inputs from a single base period prior to the target period, other possibilities exist such as using multiple years of inputs (for example, the independent variables drawn from a two-year period preceding the target year) or employing a lag between the base period and the target period.

Suppose the vector x in Equation 2 is the same for both for the concurrent and the prospective models. The predicted risk scores for the individuals will differ between the two models as the estimated regression coefficients (the vector β from Equation 2) will not be the same for the concurrent and prospective models. For example, treatment for nonchronic medical conditions such as injuries, pregnancies or acute illnesses such as bronchitis is typically confined to a single episode of care. If the condition is included in the model, a prospective weight for such a condition would be very small. Yet most often, these kinds of conditions are not included at all in prospective models, since there are no substantive costs to predict. As a practical matter, prospective models would not typically include such nonrecurring conditions. In a concurrent model, the coefficients for such conditions may be statistically significant, as the model is being used to explain costs in that same year.

Another notable difference between concurrent and prospective models is the importance of demographic inputs as explanatory variables that are mentioned in Section 2.2.3. In a prospective model, demographic variables are typically highly significant, as these models do not have information on acute conditions that materialize during the target year, or they represent latent variables of unknown conditions that exist or may occur. As such, demographic information are useful predictors of this component of health care costs. In a concurrent model, acute conditions that contribute to an individual's risk may be included in the model, and the coefficients for demographic information may not be statistically significant.

Model fit is another key distinguishing characteristic between concurrent and prospective risk scoring models. Because concurrent estimates are based on diagnosis and other information in the same time period as the expenses that are being explained by the risk scores, such models explain a far greater percentage of variance. The SOA's 2007 study "A Comparative Analysis of Claims-Based Tools for Health Risk Assessment" compared the fit of the model of a variety of concurrent and prospective models (Winkelman & Mehmud 2008). The R^2 statistic associated with the prospective models included in the study ranged from 12.9% to 21.3%. The R^2 statistics for the concurrent models were much higher and ranged from 27.4% to 49.8%.

Despite the difference in model fit, both types of models have their place in practical application. As stated before, concurrent models are most useful for explaining costs in the current period, while prospective models are more appropriate for predicting costs in a future period. In the context of program evaluation, the type of model to be selected would vary based on the type of analysis to be conducted. A concurrent model is useful in an analysis normalizing per member per month medical costs for health risk. For example, the effects of changing risk could be removed from historical PMPM medical costs to isolate other effects not related to risk changes.

A prospective model is more appropriate in a formal treatment and control approach for program evaluation. In such an application, individuals are randomly assigned to either a treatment (receiving some sort of intervention) or a

control (not receiving intervention) group. A prospective risk scoring model is then used to measure expected costs for both groups and the actual costs measured to determine the effect of the intervention.

Choosing whether to use a concurrent or prospective risk scoring model in a health insurance system depends on the nature of the risk transfer in the system. A concurrent model inherently places less risk on the issuer because unpredictable events that occur during the year are accounted for in the adjustment. For example, an adjustment is completed for a disproportionate share of pregnancies using a concurrent approach, but a prospective application does not directly acknowledge pregnancy-related risk. This risk is reflected in demographic factors and will leave the issuer at risk for any deviations from the average rate of pregnancy-related risk. We describe programs using either a concurrent or prospective risk scoring approach in Section 3.

2.2.3 Potential Explanatory Variables

The person-level inputs used by most risk scoring approaches are drawn from three primary sources: encounter records (medical claims and prescription drug records), demographic data, and other administrative or clinical data as discussed below.

Typically, the most important data in terms of distinguishing one individual's medical risk from another are drawn from encounter records, as an individual's utilization of health services varies depending on the conditions that they have, as well as the severity of the condition. A claim record will typically have one or more diagnosis codes that indicate a specific medical condition that was treated or was related to the condition being treated during the encounter. With thousands of diagnosis codes available, modelers use disease classification systems to more efficiently group codes into similar categories for inclusion in the model, such as heart disease or diabetes. One such classification system, the Diagnosis Related Groups (DRGs), initially developed in the Medicare program, is discussed in Section 3.

While diagnosis codes are essential inputs to risk scoring models, some models incorporate other logic that is designed to medically verify that a particular condition is present. One common problem with diagnosis data is that "rule-out" diagnoses are sometimes entered onto the claim form. These are diagnoses that reflect the work of a provider checking to see if a patient has a condition or not (i.e., to "rule it out"). Thus, the appearance of such a code does not necessarily indicate the presence of the disease. Pre-processing algorithms vary in their attempts to solve this problem. Most approaches exclude these rule-out diagnoses from any consideration, while others may require corroborating diagnoses from a second encounter visit or from other clinical areas, such as laboratory records or radiology reports. When the process for diagnostic inclusion in a model is made more stringent, some individuals who do not have the condition in question will not be identified, thus reducing the risk of false positives. However, with more stringent criteria for including a diagnosis for an individual, there is an increased chance of false negatives, which is not identifying someone who has the disease. Developers of risk scoring models attempt to determine an appropriate balance between these considerations.

Some risk scoring models also use procedure codes as part of their classification logic. A procedure code is a record of specific services a provider conducted during a visit. This could include a specific surgery being performed or an evaluative visit with a physician. While the inclusion of procedure codes can increase the predictive accuracy of a model, it can also introduce incentives for potential gaming of the risk score. At one extreme, a provider who is capitated using risk adjustment payments could be influenced to choose a particular procedure for treatment because of the known effect on the risk score, and thus on their income. A more common situation is that additional or potentially unnecessary codes may be entered on a claim record, which is also a concern for using diagnosis coding in a risk scoring model. If codes are entered with greater specificity, the risk scores will likely increase without any corresponding increase in actual disease burden. One recent study estimated that enrollees in risk-adjusted Medicare Advantage plans show risk scores that are 6% to 16% higher than comparable Medicare beneficiaries who are not enrolled in such plans (Geruso & Layton 2015).

Many risk scoring approaches also incorporate pharmacy data as inputs. Drug codes are mapped to clinical categories, either with unique pharmacy categories or by mapping pharmacy experience into the same categories that are also determined by diagnosis coding. The use of pharmacy data as an input brings some distinct advantages. Models incorporating pharmacy data are generally more accurate in the aggregate, because pharmacy data provide a more

specific and more recent snapshot of certain conditions. For example, an individual may be taking regular maintenance medications for a condition and that drug will be present on claim records every month, though it may appear on a medical claim only once a year when the individual visits a physician. Depending on the level of sophistication of the mapping algorithm (that is, the approach used to translate instances of pharmacy experience into aggregated condition groupings), certain strengths or combinations of medications could indicate severity differentiations that cannot be picked up by diagnosis codes alone.

Pharmacy inputs do bring added difficulties, however. In particular, available drugs and their associated codes change frequently, and managing the updates can be burdensome, often resulting in out-of-date mappings. Another difficulty of using pharmacy data as input for risk scoring is that alternate or off-label (where a drug is used to treat additional conditions other than its original intended use) uses for specific drugs sometimes emerge. This can create potential disconnects between the experience on which a model is constructed and the environment in which it is being applied. For example, the beta-blocking drug nadolol is commonly prescribed off-label for the management of migraine headaches (U.S. National Library of Medicine 2015). While a risk scoring model's mapping may view nadolol use as an indication of hypertension or other heart disease, the enrollee using the drug may not have heart problems, but actually migraine headaches. In this latter case, a misestimation of their associated risk results.

Some risk scoring models are beginning to include clinical or biometric data in their risk classification approach, as the data can provide additional detail pertaining to the severity of a condition (Iezzoni 2012). As demonstrated by the life insurance industry's underwriting practices, information such as body mass index, blood pressure, cholesterol readings or results from genetic testing provides additional insight into the riskiness of an individual. In applications where risk scoring models are being used to forecast future risk as discussed in Section 2.2.2, such data can be a measure of undiagnosed medical risk, conditions that have yet to be diagnosed, or conditions that are just starting. While important in predicting medical costs, an individual's BMI or lab results may be more useful in prospectively estimating risk than in evaluating concurrent risk, though these peripheral data points may indicate degrees of severity for certain conditions. A major limiting factor in the use of clinical and biometric data is its availability. These data are typically not transferred or available to health plans and, as such, are not broadly useful. A summary of the challenges and predictive possibilities of incorporating clinical data is provided by Parkes (2015).

Health care expenditures in a prior year are sometimes also used as inputs to the risk scoring process. The historical cost of treating recurring conditions may provide substantially more information than a simple indication of the existence of a diagnosis, as it includes some level of severity of the disease (increased expenditure is associated with higher severity). While the inclusion of prior cost in a prospective application does typically increase the predictive power of a model, the inclusion of this input variable may also introduce unintended consequences. For instance, some costs from a prior year may be associated with a condition that is nonrepeating (e.g., pregnancy and infant delivery, routine surgery or treatment of an isolated injury). The presence of such conditions influences cost in the year from which the inputs are being drawn but have little to no influence on costs in a subsequent year. Diagnosis codes for these conditions are not included in the risk scoring model for this reason. An additional concern with using prior year costs in a risk model is that it introduces a counterproductive incentive. When the intent of a risk scoring application is to reduce or manage costs, the use of prior costs as an input to a risk score creates a situation where higher costs in one year are associated with a higher risk score in the next year. This is the opposite effect to what is desired. However, purely from a predictive accuracy perspective, the use of prior costs does bring in additional information about an individual's risk that is not captured solely by diagnoses (Schone & Brown 2013).

Demographics are a key input parameter for most risk scoring models. Age and sex are critically important in prospective models when conditions are diagnosed within the year for which risk level is being evaluated. Models generally include different independent variables to account for combinations of age and sex. There might be one variable for infants, one for children and one for adults. There may be one variable based on sex, or the age categories could be separately identified by sex.

Age and sex may be useful in differentiating the risk contributed by certain conditions. Suppose the independent variables in a linear model include a list of disease conditions, as well as age and sex. A specific condition may have a different impact for one age range than for another. One way to account for this is to include an interaction term between all of the disease codes, age and sex.

For models that endeavor to explain costs using diagnoses from that same year, age and sex provide a less clear-cut advantage in terms of explaining costs. While it is generally understood that health care costs increase with age, that relationship is driven by the conditions that a person acquires over time, not strictly by the age of the individual. When an individual's full diagnostic profile is properly incorporated into a model, their age and sex become considerably less important. In other words, age and sex are proxies for complete diagnostic information, and the value of such proxies diminishes as more complete and accurate diagnostic information are made available. In some instances, demographics may not be statistically significant in a concurrent model with complete diagnostic information, a somewhat surprising outcome.

2.2.4 Estimation of the Risk Score

Once the choice of the independent variables is made as a result of the usual statistical modeling process, the risk score is estimated. The calculated risk score is a function of the independent variables and the parameter estimates of the unknown regression coefficients β. This new vector of estimates is labeled $\hat{\beta}$, with the resulting risk scores labeled $\hat{y}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \qquad (3)$$

If the $\hat{\beta}$ are unbiased for β, then taking the expected value of $\hat{y}_i$ is equal to

$$\begin{aligned} E[\hat{y}_i | x_i, \beta] &= E\left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}\right] \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\ &= E[y_i | x_i, \beta] \end{aligned} \qquad (4)$$

Thus, the average risk score, conditional on a profile of explanatory variables, is equal to the true average risk score. Note that the conclusion reached in Equation 4 does not state that the particular risk score predicted for an individual is equal to his or her observed risk score. The statement says that the average of the predicted risk scores for individuals for the values of the explanatory variables will be equal to the true average value for individuals with the same values for each of the independent variables.

As is true in general for regression models, if the variables for the true model are not as that shown in Equation 3, then there could be bias in the prediction of the risk score. Usually including too many variables (overfitting) will not result in much prediction error even if the coefficients for the variables are not statistically significant. The risk of overfitting is that collinearity can be induced among the explanatory variables and the residual standard error of the regression will be increased (Frees 2010). If the model does not include variables that are statistically significant (underfitting), then there is also bias in the predicted risk score (called omitted variable bias). Because the perfect set of explanatory variables is likely not included by any one particular risk scoring approach, any practitioner using or developing these models should be aware of the consequences of the associated bias.

2.2.5 Recalibration of Model Coefficients

Vendors of risk scoring models typically include a set of coefficients for use in determining risk scores as described above. There are instances when it may be appropriate for an actuary to estimate a set of alternate weights through a process called "recalibration." Actuarial Standard of Practice 45 states that "recalibration is often used to make the risk adjustment (or risk scoring) model more specific to the population, data, and other characteristics of the project for which it is being used" (Actuarial Standards Board 2012). *Weights* are the term commonly used to refer to the estimated coefficients in a risk scoring model. For example, a variables representing a disease condition will have an associated weight that indicates the incremental effect on the risk score.

Recalibration can properly align the population and outcome variables when the original development of a risk scoring model's estimated coefficients was performed using a population that is substantively different. For example, researchers often employ risk scoring models that were developed on Medicare (CMS-HCC) or Medicaid (CDPS) populations. While these models have the advantage of being free and completely transparent, they are designed and

validated for specific populations that may not be appropriate for a commercial insured population. In these instances, recalibration is recommended for optimal results. A difference in outcome variables may also call for recalibration. For example, a model may be designed to predict total allowed costs, but the specific measure of interest is paid amounts associated with a particular benefit design. In this scenario, the original weights would represent the risk in terms of total allowed costs, but the risk associated with the specific benefit design would differ and would require new coefficients.

A full recalibration of a model would entail respecification of the regression equations that were initially employed in the development of the model. This is not always practical or possible. One barrier to this approach is that the model in use may not be adequately transparent to allow for a full recalibration. To fully recalibrate, one would need to know all of the independent variables that were used in the model and know how they were developed. Additionally, the logic behind any interaction effects or hierarchies (for example, combinations of conditions or instances where the presence of one condition precludes the inclusion of a second) would also need to be implemented in the recalibration, and this information may not be released by the model developers. Another practical difficulty that can be encountered is the availability of an adequate sample size for the recalibration. Some rare conditions may be observed with minimal frequency and require a larger population than is available for the development of statistically significant coefficients.

Actuaries have developed approaches to recalibrating model coefficients without performing a full recalibration. In the SOA's 2007 study by Winkelman and Mehmud, the authors performed regressions where the dependent variable was the residual from each model, or the difference between the actual value and the predicted value (Winkelman & Mehmud 2007). The independent variables included in the regression were the condition categories that were visible to the study authors for each model. By regressing on the residuals rather than the actual cost, any variables or interactions that are unknown to the modeler are still implicitly retained within the recalibrated model. A more recent article proposes a similar approach but employs a penalized ridge regression to indirectly encourage the model to prefer the original risk scores unless sufficient data are present to override with new coefficients (Parkes & Armstrong 2015). An additional study has shown that with a sufficient sample size and with full model transparency, the various approaches to risk model recalibration converge to the same result (Hileman 2015).

2.2.6 Handling of Enrollees with Incomplete Enrollment Periods

Most risk scoring approaches are designed and specified for individuals with complete data, usually 12 months of continuous enrollment in a health plan. However, in practical application, many individuals are not enrolled in a plan for a full 12 months. An individual's birth or death is one example of why someone would have less than 12 months of data, but more commonly this is due to enrolling or disenrolling in a health plan during the year. Risk scores are still needed for those without complete data to complete the risk adjustment calculations for a transfer payment or other application. Thus, adjustments are needed in the method to estimate their risk score. These adjustments can take several different forms and are discussed in the context of specific risk scoring models in Section 3.

The degree to which incomplete data affect the accuracy of a predicted risk score varies by both the length of the missing enrollment period and an individual's medical conditions. Clearly, a shorter period of enrollment provides the model with less opportunity to record relevant medical conditions. Individuals seeing a provider more frequently may have less underreporting than those who see a provider less frequently. Also, models that incorporate prescription drug data will record chronic conditions when individuals receive their drugs, even though they may not be treated by their provider within the observation period.

2.2.7 Limitations of Estimated Risk Scores

Due to the random nature of health care events with their chance of occurring, the severity of the event and the chance of the individual receiving care, predicting risk scores (or medical expenditures, defined earlier) is a difficult endeavor. Even the most acclaimed risk scoring methodologies are not highly accurate when measured by traditional statistical measures. In a 2007 SOA study, most prospective models resulted in a R^2 between 20% and 30%, meaning that the majority of variation in medical expenditures was left unexplained by the models (Winkelman & Mehmud 2007). Because of the randomness of health care events, variation will remain that cannot be explained by any risk scoring methodology, even with perfect information about an individual's medical conditions. A small number of

individuals have very large total medical expenditures, while many individuals have little or no expenditures in a year. There is variability within an individual from one year to the next in their need for medical care. Other factors that influence this variability include the individual preference for different medical treatments or settings, given the same medical diagnoses. Individual medical practitioners have different habits or preferences in treating the same diagnostic conditions, including the tests chosen, techniques pursued or drugs prescribed. Geographic differences in practice patterns or unit costs for the same procedures cause additional variation. Variation in access to care can also lead to differences in medical expense that cannot be explained using diagnostic and pharmacy data.

Additional variation is introduced by incomplete or inaccurate information about the diagnostic profile of an individual that can result from bad data or incorrect coding by the providers or a limitation of the coding schema (for example, a database that is designed to hold only five of the possible diagnoses associated with a claim record). A diagnosis code does not convey complete information of the severity of a condition or its likely costs of care.

Recall that for linear models, errors are normally distributed and are constant over all values of y given the explanatory variables. The predicted risk score can be too high for low expenditure values and too low for high expenditure values. For example, an individual with zero medical costs in a given year will typically have a prospective risk score than is greater than zero. Conversely, an individual with very high claims costs is likely to be underpredicted by a risk model. One reason is that normally distributed errors in the model do not have heavy tails (that is, the actual distribution of health care expenditures will have more frequent extreme high outliers than the distribution of predicted risk scores). Also, the model assumes that the variance is constant given the explanatory variables. It could be that the independent variables used in a risk model (or available) are able to explain only some of the variation in medical costs. Two individuals with the exact same diagnostic and demographic profile may have very different total expenditures, due to personal preferences, the course of treatment, the cost of specific providers utilized or random chance (for example, one individual has a complication during surgery that results in greatly increased costs and the other does not). In the 2007 SOA comparison study, individuals in the 99th–100th percentile of expenditures were underpredicted on average by about 75%. Individuals in the 20th–40th percentile of expenditures, only slightly below the median, were overpredicted by about a factor of five. The estimated risk scores for individuals very close to zero expenditures were even more overestimated, by factors of 50 or more (Winkelman & Mehmud 2007).

The predicted risk score, as detailed in Section 2.2.4, is a function of the independent variables and is an average risk score reflecting the profile of explanatory variables in the model. The impact of explanatory variables not included in the model, called *omitted variables*, would be absorbed by the error term or some of the explanatory variables included in the model if the omitted variables were correlated with them. If the *true* model was really

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_{p+1} x_{ip+1} + \varepsilon_i \quad (5)$$

then the predicted value of y_i remains the same as in Equation 3 as the omitted variable is not included, yet the *true* expectation is

$$E[y_i | x_i, \beta] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_{p+1} x_{ip+1} \quad (6)$$

Thus, there would be a difference of $\beta_{p+1} x_{ip+1}$ between the predicted and true risk scores for individual i. In general, if these omitted variables were significant, then the predicted risk score is biased because of the omitted variables. This is the danger of underfitting a model.

Risk scores can change over time without any underlying change in the underlying risk represented by a population. One reason is known as coding drift that was originally discussed when Medicare introduced the Prospective Payment System and the DRGs (Carter et al. 1990). Further discussion of DRGs is found in Section 3.1.1. Coding drift is due to an increased number of coding slots availability on a medical record. An artifact of an increased reliance on risk scoring for use in provider payment is that there is an increasing incentive for providers to code diagnoses more completely. Coding practices have also changed due to improvements in both technology (for example, electronic medical records systems that can supplement the manual coding process) and data storage (older databases may simply not have held enough positions to record all of the coded diagnoses) (Iezzoni & Moskowitz 1986; Jencks et al. 1988). On the other hand, risk scores may decrease over time if the underlying models and mappings are not maintained to keep up with

changes in health care practice. For example, if a model includes prescription drugs as covariates, these must be regularly updated as new drugs enter the marketplace. These mappings must be maintained so that individuals using new drugs will be treated similarly to individuals using older drugs.

2.3 Model Selection

As mentioned in Section 2.1, risk scoring models have existed for 40 years and were developed originally for specific purposes, such as the analysis of hospital mortality statistics. Since the introduction of the DRGs in the Medicare Prospective Payment System in the early 1980s, discussed in more detail in Section 3.1.1, an increasing number of risk scoring models have been developed and for commercial purposes. As mentioned in Section 2.2.3, the Society of Actuaries has conducted several studies comparing different commercial products that are for the purposes of risk scoring (Dunn et al. 1996; Cumming et al. 2002; Winkelman & Mehmud 2007). Users wanting to apply a risk scoring model are faced with many options. The purpose of this section is to provide a guide to describe the various factors that influence the selection of a model for a particular application.

The objective of these comparisons is not to produce a ranking or to state that one model is the best model that should be used in every situation. To decide between models requires an educated understanding of the eight Risk Principles listed above.[1] The following paragraphs discuss these Risk Principles and their impact on the selection of a risk scoring model by a user.

The purpose of the development of the original model and the user's purpose need to be aligned. Using a model that predicts 30-day mortality would be not be appropriate for a study to analyze expenditures or risk scores.

As discussed in Section 2.2.5, ideally the model preferred is one that was developed using a population that is similar to the user's application, as the regression coefficients (the vector of β) used to determine the risk score are a function of the population used to develop them. Otherwise some method of recalibration is needed to adjust the coefficients to adjust to the user's population.

Knowing whether a model is applied on a concurrent or prospective basis, as discussed in Section 2.2.2, is of interest to the user. The risk scoring model used for Medicare Advantage, discussed in Section 3.1, is designed using a prospective framework. As discussed in Section 3.4, the choice of a concurrent approach for the ACA's commercial marketplaces was in part due to concerns about data being available from a prior year for a newly insured population or for a highly transient population who frequently shift from one carrier to another. From a policy perspective, the choice between a concurrent or prospective model may be guided by the degree to which risk is being transferred to a risk-adjusted capitated entity. For example, in a prospective setting, carriers would be at full risk for one-time occurrences such as injuries, pregnancies and acute isolated illnesses. However, the risk associated with these conditions would be potentially built in to a concurrent application.

While data sources need not be identical, the data required for input to any risk scoring model need to be available to calculate the risk scores. As discussed in Section 2.2.3, models may utilize some combination of pharmacy data and medical diagnosis data. If a study to be conducted requires explanatory variables that are not used in the particular risk model, the predicted risk score would not result in an accurate prediction as a potentially significant variable is omitted. If there was a particular need for accuracy by demographic category or medical conditions, risk scoring models that include the appropriate explanatory variables might be more appropriate to the user's purposes.

The model development approach behind a particular risk scoring approach can inform the appropriate application. Some models were developed from more of a clinical perspective and may have more ancillary uses beyond the generation of risk scores. A model developed from a purely statistical framework may have better predictive performance but fewer applications in the clinical environment.

[1] The principles are: (1) purpose, (2) population of interest, (3) time frame, (4) data source, (5) variables/factors included, (6) model development approach (whether statistical, clinical or both), (7) acceptable level of accuracy of results, and (8) implementation approach.

Users of risk scoring models will want to understand the degree to which various models can accurately predict risk levels for individuals and for groups. As discussed in Section 2.2.4, the predicted risk score is a function of the estimated regression coefficients and the covariates in the model. As noted there, the individual predicted risk score, an estimate of the average for the given covariates, will differ from the observed risk score. However, for a group of similar individuals, the predicted risk score will be more reflective of the *true* mean due to the larger sample size. This statistical property is utilized in actuarial work, as the risk adjustment process is completed at a group level and not at a member level. Other measures of model fit that may be used to compare models on a more global level include R^2, AIC and sum of squared prediction errors.

Implementation of the risk scoring model, the final risk principle, consists of three elements: flexibility, transparency and financial cost. Flexibility would exist if the model could be recalibrated or whether the condition mappings can be adjusted. Transparency aids in the communication and acceptance of a risk scoring model. Documentation that shows the mapping of diagnosis or pharmacy codes to groups and disclosure of the regression coefficients aids in the transparency. As a rule of thumb, a model may be considered transparent if sufficient information is accessible to the user such that the user can calculate an individual risk score manually and match the results of the model. The financial cost of a model can include substantial licensing costs as well as operational implementation costs. There are also potential extensive computing costs, such as requiring other software, storage or computer cycle time. Finally, there are training costs to implement a model within an organization.

Section 3: Applications of Risk Scoring and Adjustment

This section provides an overview of the manner in which risk scoring techniques have been incorporated into the health care financing environment. The use of risk scoring in government entitlement programs (Medicare and Medicaid) is explained, as well as the mechanism for ensuring appropriate financial transfers between commercial health plans in the post-ACA marketplace.

3.1 Medicare

The Medicare program was established in 1965 as Title XVIII of the Social Security Act and was effective July 1, 1966, establishing a health care program for persons aged 65 and older (CMS 2013). The original reimbursement approach was fee-for-service. As discussed below, risk scoring for the Medicare managed care services was considered in the 1990s but not introduced until 2004. The 2003 Medicare Prescription Drug, Improvement, and Modernization Act created Medicare Part D, a voluntary drug benefit program for those with fee-for-service plans (Robst et al. 2007).

The discussion below begins with a summary of the Diagnosis-Related Groups (DRGs) that were used originally by Medicare as a way to reimburse hospitals for inpatient stays, but are now used by both public and private insurers. While not technically a "risk adjustor," the development of the DRGs serves as a foundation in the ways that risk adjustors are being developed today as this was an early instance of the grouping of diagnoses into units of more homogeneous risk units. This discussion is followed up with a summary of risk adjustment in Medicare managed care and in Medicare Part D.

3.1.1 Diagnosis-Related Groups (DRGs)

One widely known statistically based protocol is the Diagnosis-Related Groups (DRGs), introduced in 1982 for use in reimbursing hospitals for care for patients in Medicare. The development of the DRGs began in the late 1960s as a tool to measure and evaluate hospital treatment by creating a manageable number of stable, clinically meaningful groups to establish prices, estimate costs and evaluate quality (Fetter et al. 1991). The DRGs were developed based on a combination of physician expertise and statistical methods. Fundamental to the development of the DRGs is the ICD-9-CM classification system. The statistical approach to develop the DRGs incorporated techniques from survey sampling

(stratified sampling) to select the data, and machine learning techniques (recursive partitioning) to create meaningful groups. The scheme to develop the DRGs had five characteristics (Fetter et al. 1980):

1. Must be interpretable medically, with subclasses of patients from homogeneous diagnostic categories.

2. Individual classes defined on variables commonly available on hospital abstracts and relevant to output utilization, pertaining to either the condition of the patient or the treatment process.

3. Must be a manageable number of classes, preferably in the hundreds instead of thousands (ICD codes), that are mutually exclusive and exhaustive.

4. Classes contain patients with similar expected measures of output utilization.

5. Class definitions must be comparable across the different coding schemes, for example a revision of the ICD.

An interactive process called AUTOGRP, a modification of the Automated Interaction Detector, was used to create a partition of the data (Sonquist & Morgan 1964). The data were separated into groups that minimized the unexplained variance of the dependent variable, length of stay. Certain patient attributes were used to explain the variation in resource consumption based on length of stay. The independent variables included operational procedure, principal diagnosis, age of patient, sex of patient, presence or absence of specific complications and/or comorbidities, and discharge status. Other variables, such as identity of payer, number of comorbidities or complications and number or type of diagnostic procedures, and type of admission, were considered for use but rejected (Mills et al. 1976). The development of the DRGs—in particular, the grouping of diagnoses to produce groupings of hospital stays that represent homogeneous risk units—served as a blueprint for later development of diagnosis-based risk scoring methodologies.

3.1.2 Medicare Managed Care

In 1973 the HMO Act established health maintenance organizations (HMOs), which were authorized to provide services to Medicare recipients (Weissman et al. 2005; CMS 2013). Reasonable costs were reimbursed through this program. Through the Tax Equity and Fiscal Responsibility Act of 1982 (TEFRA), a new approach to reimburse for services was introduced called Adjusted Average per Capita Cost (AAPCC) (Weissman et al. 2005). Medicare HMO services were reimbursed at 95% of the fee-for-service rate, along with adjustments for county demographics, age, sex, Medicaid enrollment, working status and institutional status. Studies in the 1980s showed that the AAPCC methodology overcompensated HMOs for services provided, as enrollees in HMOs were healthier and used fewer services than enrollees in a fee-for-service arrangement (Greenwald et al. 1998; Weissman et al. 2005). A study by Brown et al. (1993) showed that HMO costs were almost 11% less than those in fee-for-service plans; thus the 5% discount was not sufficient to account for the differences in costs between managed care and fee-for-service patients.

The deficiency of the AAPCC methodology led to the development of health risk adjustors in the early 1990s. The Balanced Budget Act of 1997 (BBA) included provision for the use of risk scoring models in Medicare. The coefficient of determination statistic, R^2, was first used to compare models; however, with the low values for R^2, proponents of risk scoring decided to use predictive ratios (predicted expenditures vs. actual expenditures) for groups of people as these ratios were more persuasive. These predictive ratios were calculated for various subgroups, such as the lowest cost or highest cost quintiles, or various specific diseases such as acute myocardial infarction or hip fracture. The use of predictive ratios was helpful to persuade policymakers on both sides of the aisle to vote for the 1997 BBA. However, inclusion of risk scoring in Medicare managed care did not occur soon after passage of the BBA, as issues arose as part of its implementation. These issues can be summarized as (1) model selection and the data burden, (2) time lag of claim data and (3) the effect of risk adjustment on Medicare costs (Weissman et al. 2005).

The two initial leading models considered were those based on the ACGs (Ambulatory Care Groups or Adjusted Clinical Groups) and the DCGs (Diagnostic Cost Groups). The ACGs were originally developed for outpatient use, while the DCGs were originally developed for inpatient use. A long list of literature supports the development and application of each of these models (Starfield et al. 1991; Weiner et al. 1991; Ellis & Ash 1995; Ellis et al. 1996). Philosophically, the

idea of including encounter-based models was appealing to those in government, as the data were more comprehensive; however, the data burden on health plans would be large. The initial choice was to implement the DCG model renamed the PIP-DCG (Principal In-Patient Diagnostic Cost Group) with an eye toward migrating to an encounter-based system in the future (Pope et al. 2000). One concern related to the PIP-DCG model was that it was based on only inpatient costs, which rewarded health plans for hospitalizing patients rather than managing their care (Weissman et al. 2005).[2]

Medicare wanted to use a prospective approach to risk scoring. As discussed in Section 2.2, a prospective model uses data from the current year to predict the next year's expenditures. From a timing standpoint, health plans would need access to data for the entire year before setting their rates for the upcoming year. This was not feasible. An alternative approach was adopted to collect data from the first six months of the current year and the last six months of the prior year to allow sufficient time for rates to be set.

Policymakers determined that controlling spending under Medicare was critical to help balance the federal budget as part of the BBA of 1997. As the Congressional Budget Office initially stated that risk scoring would be budget neutral, health plans were opposed to the federal government keeping the savings estimated at $10–11 billion over five years (Weissman et al. 2005).

In 2002, Medicare shifted to the DCG-HCC (Hierarchical Coexisting Conditions) model that used 61 disease groups and used both inpatient and outpatient data. In 2004 Medicare implemented risk adjustment on a budget-neutral basis. The DCG-HCC ultimately shifted to the CMS-HCC model, which simplified data collection. The CMS-HCC model was built on 10 principles, similar to that proposed by Feinstein and used in the development of DRGs (Feinstein 1967). Instead of predicting length of stay as for the DRGs, the CMS-HCC model predicts health care expenditures. Rules were created to define minimum sample size for the disease categories. The rules did not reward or penalize health plans for coding extra conditions as it used the highest ranking condition in a particular hierarchy. High-cost diagnoses that were not prevalent were excluded. In the end, there were 70 categories representing 3,000 of the 15,000 ICD-9-CM codes. For this model, data from 1999 were used to predict claims costs in 2000. The data lag as discussed above would be eliminated because rates would be established as of January 1 and then finalized on June 30, with any necessary retroactive adjustment. The model inputs included age, sex, an indicator of Medicaid eligibility, an indicator for original Medicare entitlement through disability, 70 HCC categories, interactions of Medicaid with sex and disabled status, and six disease interactions (Ellis & Ash 1995; Ellis et al. 1996; Pope et al. 2004, 2011).

Even with all of the potential impacts, risk scoring was introduced into Medicare managed care because it benefited a wide variety of stakeholders (Weissman et al. 2005). Risk scoring ended the unfair advantage that managed care plans had relative to fee-for-service plans.

3.2 Medicaid

The Medicaid programs were established in 1965 as Title XIX of the Social Security Act and became effective July 1, 1966 (CMS 2013). As the programs are jointly administered by the federal and state governments, the program is highly complex and varies greatly from one state to the next with respect to the groups of individuals covered, the scope of services included and the mechanisms (including risk-based capitation) employed. Thus, it is beyond the scope of this paper to elaborate on the specifics of how risk adjustment is used within each state's Medicaid program.

However, risk scoring is an important element of rate setting within many Medicaid managed care programs (Center for Health Program Development & Management, & Actuarial Research Corporation 2003). Risk scoring models are used in Medicaid to properly match capitation payments with an estimate of their financial risk as measured by the risk score. Medicaid programs must evaluate several special considerations when selecting an approach to risk adjustment. First, unlike commercial or Medicare plans, Medicaid programs and their managed care contractors can experience very high turnover even within the year, as individuals can frequently enter in and exit out of Medicaid eligibility. Because risk scoring models are informed by the available diagnostic history of each individual, the lack of visibility

[2] There was an exception for congestive heart failure, where if evidence-based criteria were followed, then a bonus payment would be made.

resulting from this higher turnover can result in poorer predictive performance. Some Medicaid programs address this issue by basing risk-adjusted payments on the cohort enrolled to the plan in the most recent available period, rather than allowing the risk score to "follow" the individual from plan to plan. This approach implicitly assumes that a plan will attract the same type of enrollees from one year to the next. A disadvantage of this approach is that when the demographics/severity of enrollment changes, there will be a lag in acknowledgment of these changes in capitation payments.

Another issue that is particularly important within the Medicaid risk adjustment environment is that there are often very distinct categories of individuals receiving benefits under a single Medicaid program. Two of the most common categories are those individuals receiving Supplemental Security Income (SSI) and the Temporary Assistance to Needy Families (TANF) population. The SSI beneficiaries tend to be longer-term Medicaid beneficiaries and can include high concentrations of individuals with chronic conditions, or individuals who are blind or have developmental disabilities. The TANF population, however, exhibits much higher turnover and tends to have less variation in their need for health care services as many such beneficiaries are healthy children. Other populations that may be considered for risk adjustment include the SOBRA population, consisting of certain low-income pregnant women, and the Medical Assistance only groups, who qualify for Medicaid by virtue of particular health conditions not related to income (Center for Health Program Development & Management, & Actuarial Research Corporation 2003).

The issue of model recalibration is particularly important for the Medicaid population as well. Although there exists a free, publicly available, risk scoring model that is specified on a Medicaid population (the Chronic Illness and Disability Payment System [CDPS]) (Kronick et al. 1995), its weights do not necessarily reflect the enrollment mix and coverage parameters in a given U.S. state. Many programs carve out certain services, such as those associated with pregnancy and childbirth or mental health services. That is, such carved-out services are not included in the capitation payment for which the risk adjustment approach is being applied. In the event of carved-out services, any associated expense is excluded from the development of the model weights to be implemented and can be financed through the use of "kick" payments that cover these costs on a per-episode basis without subjecting them to risk adjustment (Winkelman & Damler 2008).

3.3 Medicare Part D

Prior to the enactment of the Medicare Modernization Act of 2003, Medicare did not provide coverage for outpatient prescription drugs. The Part D program closed that coverage gap by establishing outpatient drug coverage through commercially provided prescription drug plans. These plans operate in a similar manner to Medicare Advantage plans, with an annual open enrollment period during which beneficiaries can select a plan. Some Part D plans are integrated within Medicare Advantage plans, while others are designed to stand alone. As in Medicare Advantage, a risk scoring mechanism was introduced to normalize payment rates among plans that attract different levels of pharmacy expense risk. The Part D HCC model (RxHCC) was developed on a similar conceptual basis as the CMS-HCC model, with medical diagnoses (not pharmacy encounter data) mapping to up to 84 diagnostic condition categories (HCCs). Each HCC has an associated weight that represents the incremental risk posed by the presence of that condition to a Part D plan. While the RxHCC model explains an even smaller proportion of risk than the Medicare Advantage CMS-HCC model, the mechanism plays an important role in reducing the risk of adverse selection for Part D market participants (Hsu et al. 2009).

3.4 Affordable Care Act (ACA)

Before 2014, the role of risk scoring within commercial health insurance was generally limited to the realm of clinical informatics and studies concerning the health of the insured population. The introduction of rating reforms mandated by the Affordable Care Act (ACA) has vaulted risk scoring to a central position in commercial health insurance. Prior to the ACA's reforms, in most states, health insurance premium rates were closely matched to the risk level of the insured individual or small group through medical underwriting. Carriers used tools such as detailed medical questionnaires, medical examinations and analysis of incurred claim levels to determine the appropriate rate for each new or renewing individual or group.

Under the ACA's rules, carriers are now much more limited in the extent to which premiums can vary from one person to the next. Premium variation in the individual and small group markets is permitted by only age, geographic region, family composition and tobacco use status. Age rating is also limited to a ratio of 3:1 from the highest-rated age to the lowest-rated age (natural claims experience typically suggests a much steeper curve). Notably absent from the list of permissible rating factors are sex and individual health status. These mandated changes in the rating structure introduced new subsidies into health insurance pricing: younger enrollees now subsidize older enrollees, male enrollees subsidize female enrollees, and healthier enrollees subsidize sicker enrollees. While these rating reforms help achieve one of the ACA's stated goals of making health insurance more uniformly affordable, they open the door to a much greater risk to the carriers of adverse risk selection. In other words, when rates could more closely match the risk level of the insureds, carriers were able to charge appropriately higher premiums for higher risks. Under the ACA's rating reforms, a carrier insuring a high-risk individual would receive the same premium as for a lower risk individual, provided that the individuals were the same age.

Along with these rating reforms, the ACA did establish three key programs designed to mitigate risk to carriers, known as the "three Rs": reinsurance, risk corridors and risk adjustment. The first two programs, reinsurance and risk corridors, are temporary programs designed to help minimize risk and stabilize premiums over the first three years of the rating reform implementation. The reinsurance program established a pool from which certain high-cost individuals' claims would be paid. This pool is funded through mandatory contributions from all health plans, even those that do not participate in the reinsurance program. The risk corridor program provides aggregate stop-loss protection to carriers by insuring a portion of the difference between the carrier's actual cost level and an expected target level.

While the reinsurance and risk corridor programs will be phased out by the 2017 plan year, the risk adjustment program is a permanent—and critical—piece of the commercial health insurance landscape. Risk adjustment in the commercial marketplace is designed to ensure that carriers attracting high-risk individuals will be able to compete on a level basis with carriers that attract lower risk individuals. This allows for "the focus of plan competition to shift from risk selection to quality, efficiency, and value" (Kautter et al. 2014b). There are two elements to the risk adjustment program: first, risk scores are calculated for each enrolled individual to determine the relative risk of each carrier in a risk pool; second, a financial transfer is determined based on a comparison of the average risk level across each risk pool by state. Within each state, there are three distinct risk pools: individual, small group and catastrophic.

For the purposes of ACA risk adjustment, risk scores are calculated using a new risk scoring methodology developed by CMS and called the HHS Hierarchical Condition Categories (HHS-HCC) model. This is distinct from the structurally similar CMS-HCC model used to risk adjust payments within the Medicare Advantage program, discussed in Section 3.1. The HHS-HCC model uses diagnoses and demographics to assign a risk score to each individual. Each individual's diagnoses are mapped to one of more than 100 condition categories. Additional logic is then applied to incorporate hierarchies into the mappings to ensure that an individual is classified into only the highest qualified cost category within a family of conditions. Interaction terms are also included to provide additional risk weight for individuals with certain combinations of conditions.

Several unique features to the HHS-HCC approach were dictated by the structure of the post-ACA commercial marketplace. First, the model is a concurrent risk scoring model, meaning that diagnoses in a year are used to predict expenditures in that same year. This contrasts with Medicare's CMS-HCC model, which is a prospective application. The concurrent approach was chosen mostly due to concerns that data would not be available for the prior period for many plan enrollees. Particularly in the early years after the implementation of the ACA, claims data for individuals in the year prior to coverage were not expected to be generally available. Thus, if a prospective approach had been chosen, the risk scores for the most-uncertain newly insured population would be based solely on demographics. Because significant annual turnover in the insured market was anticipated, the problem of data unavailability is not limited to the implementation period. A concurrent approach, because it is based on the claims experienced during the adjustment year, allows for greater alignment between the risk posed by the enrolled population and the mechanism used to mitigate that risk.

A second feature unique to the HHS-HCC model is the use of three distinct sets of model weights for three age groups: infants (age 0 and 1), children (age 2–20) and adults (age 21 and over). The use of three age groups was done out of

consideration for the very different risk profiles across these three subgroups, and across and within each of the condition categories. For example, since a condition category may include a large number of diagnoses, the distribution of the diagnoses within each category may be quite different for children versus an adult population. The separation of these three demographic groups into different sets of risk weights permits a more precise attribution of the risk posed by each.

Finally, the HHS-HCC model is designed to predict five different target variables. While most models are constructed to predict either total allowed cost or total plan liability, the focus on risk adjusting the commercial marketplace necessitated a more carefully measured approach. The HHS-HCC model produces risk scores for each of the five ACA "metal levels": platinum, gold, silver, bronze and catastrophic. Each of these metal levels represents a different level of plan actuarial value or benefit richness to reflect the relative risk posed by an individual differs by metal level. This difference can be illustrated by a comparison of the risk associated with a high-deductible plan versus a plan with first dollar coinsurance. In the high-deductible plan, the actual risk remains at zero until the deductible is passed, then risk increases quickly. In a first dollar coinsurance plan, the risk score will increase in a more linear fashion.

The risk scoring model feeds directly into the second key component of the ACA's risk adjustment program, the risk transfer formula. The transfer formula is the mechanism by which funds are transferred from carriers who have attracted lower risk enrollees to those who have attracted higher risk enrollees. The transfer formula normalizes premiums for the allowable rating factors and then compares each plan's average risk-adjusted premium to its actual premium. The difference between these two values represents the shortfall or excess that must be made up by the risk adjustment transfer. Additionally, the transfer payments are ensured to net to zero across all plans in a market. On June 30, 2015, HHS released a report detailing the preliminary risk adjustment transfer payments for each plan and market. In total, 758 issuers participated in the first risk adjustment transfer program. Among individual plans, the transfers totaled 10% of premiums. For more-uncertain and lower premium catastrophic plans, the transfers totaled 21% of premiums. The small group market, which saw fewer new entrants in 2014, transferred 6% of premium in the first risk adjustment transfer.

3.5 Primary Care

As medical providers have moved toward assuming more financial risk for patient health care, risk scoring has assumed an important role in that arena. As with general medical expenses, the need for primary care services varies greatly from one individual to the next. Risk scoring can be important for comparisons across physicians, for evaluating outcomes-based quality measurements, or for measuring a population over time. Most current risk scoring methodologies are not particularly well-suited for specific assessment of primary care risk. First, most risk scoring models estimate the risk of total health care expenses, not costs specific to one venue of care. The relative amount of primary care needed is not directly correlated to total health care expenditures. A second complication is that much of the burden of primary care services is associated with the evaluation of conditions that may not yet be attributable to a specific diagnosis. This sort of risk does not correspond as closely to recorded diagnoses as the costs associated with the eventual treatment of those conditions (Rosen et al. 2003).

One practical application of risk adjustment in primary care is found in the Medicare Shared Savings Program (MSSP). The MSSP was established within the ACA as a program to encourage the development of Accountable Care Organizations (ACOs) serving the Medicare beneficiary population. Through the MSSP, participating ACOs are given the opportunity to receive a portion of any financial savings that are realized through the ACO's care interventions. The CMS-HCC (Medicare) risk scoring model is used to estimate a risk-neutral benchmark cost per person and the trend in health risk over the baseline period. This trend is then applied to the performance year costs in an attempt to measure the effects only of the ACO's interventions and not of underlying drift in risk levels.

3.6 Summary of Available Models

The tables below, from Schone and Brown (2013), provide a brief summary of risk scoring models that are currently used in the market. These tables summarize the manner in which each model uses diagnosis and treatment-based (pharmacy and procedures) inputs, the covered populations for which the models are tailored, the data used to develop the models, and a brief description of the modeling algorithm. An upcoming Society of Actuaries study will

provide a more thorough discussion of these models and their performance, as well as some more recent entrants to the market (Hileman & Steele 2016).

Table 1: From Table 1 in Risk Adjustment: What Is the Current State of the Art, and How Can It Be Improved?

System	Diagnosis Role	Treatment Role	Population	Data Sources	Algorithm
Ambulatory Care Groups [ACGs]	ICD-9 diagnosis codes used to classify beneficiaries	None	General	All claims for services	Mutually exclusive groupings of diagnoses based on clinical judgment and resource implications
Chronic Disability Payment System [CDPS]	ICD-9 diagnosis codes used to classify beneficiaries	None	Medicaid population—adult and child versions. Severity categories are based on resource use.	All claims for services	Groupings of diagnoses based on clinical judgment and resource implications. Beneficiaries may be assigned to multiple categories
DxCG DCGs	ICD-9 diagnosis codes used to classify beneficiaries	None	General	Separate inpatient and all-source versions	Categories are defined on the basis of clinically coherent diagnosis groups, hierarchically combined into HCCs. Individuals may have multiple HCCs.
Impact Pro	ICD-9 diagnosis codes used to define episodes	Procedures used to define episodes	General	All claims for services, drug claims	Episodes defined on the basis of diagnosis, procedure and drug data; each member may have episodes falling into multiple categories
DxCG RxGroups	Drug therapy category assignment may include diagnosis codes	Drug therapy categories from Rx claims used to classify beneficiaries	General	Prescription drug claims, may include claims for services	Drug therapy categories can be assigned to one or more categories
Medicaid RX	Diagnostic groups based on information from prescriptions	Prescriptions used to identify diagnoses	Medicaid population—adult and child versions	Prescription drug claims	Prescription drugs mapped to medical condition categories. Cost predicted based on medical condition and age/gender categories.

Ingenix PRG	Diagnostic groups based on information from prescriptions	Prescriptions used to identify diagnoses	Large managed care population, calibrated by enrollment period	Prescription drug claims	Groupings of prescription drugs mapped to diagnostic categories. Patient may be assigned to multiple categories.
Clinical Risk Groups	ICD-9 diagnosis codes	Included in category definition	General	Claims for services, may include prescription drugs	Mutually exclusive categories based on diagnostic and procedural criteria
Ingenix ERG	ICD-9 diagnosis codes and procedure codes used to define episodes	Part of episode definition	General	All claims for services, drug claims	All treatment information used in episode definition

Table 2: From Appendix III in Risk Adjustment: What Is the Current State of the Art, and How Can It Be Improved?

Risk-Scoring Application	Benefit Variability	Underwriting Permitted	Choice of Plan	Role of Risk Scoring
Medicare Advantage	Minimum, may add at subsidized price, HMO, PPO, FFS permitted	No	Annual choice	Setting benchmark costs for health plans in relation to government option, compensating plans for variation in beneficiary selection due to benefit design, regional variations
Medicare Part D	Actuarially equivalent, regulated formulary	No	Annual choice	Compensate plans for predictable variations in drug expenditures
Medicaid	Minimum required	Underwriting differs by state, usually varies by age and sex, set by competitive bidding, or regulated rates	Mandate, with default assignment, annual choice	Compensate plans for variations in expenditures not captured by rating cells. Selection may occur based on supplementary benefits, geographic variation, quality, network characteristics.
ACA Exchange	Benefit tiers, organization may vary	Rate bands, by age	Mandate enforced by tax	Equalize payments and costs across plans. Compensate plans for variation due to demographic factors, selection.

(Schone & Brown 2013)

Section 4: Overall Summary

Although the motivation for this paper is to provide an understanding of risk scoring in the Affordable Care Act, the use of risk scoring has been used in other insurance programs as well as in assessing programs or outcomes. As health care costs have continued to escalate over the past decades, tools that can be used to predict, explain or understand these costs have become correspondingly more important. As shown, these tools are not solely used by actuaries and underwriters, but have been widely used and understood by professionals throughout the health care system.

Due to this increased ubiquity of risk scoring, it is essential that practitioners understand the complex workings of these models and the associated uncertainties and dangers. We have provided a foundation to lay out these details, with the theoretical underpinnings of risk scoring and questions to ask when assessing different models, to users of risk scoring models.

References

Actuarial Standards Board. (2012, January). *Actuarial Standard of Practice No. 45*. Retrieved from http://www.actuarialstandardsboard.org/asops/use-health-status-based-risk-adjustment-methodologies/

Anderson, G. F., Steinberg, E. P., Powe, N. R., Antebi, S., Whittle, J., Horn, S., & Herbert, R. (1990). Setting payment rates for capitated systems: A comparison of various alternatives. *Inquiry*, 27(3), 225–233.

Ash, A., Porell, F., Gruenberg, L., Sawitz, E., & Beiser, A. (1986). *An Analysis of Alternative AAPCC Models Using Data from the Medicare History File.* Health Policy Center, Brandeis University.

Ash, A. S., Porell, F. W., Gruenberg, L., Sawitz, E., & Beiser, A. (1990). Adjusting Medicare capitation payments using prior hospitalization data.

Barnard, C., & Esmond, T. (1981). DRG-based reimbursement: The use of concurrent and retrospective clinical data. *Medical Care*, 19(11),1071–1082.

Blumenthal, D., Weissman, J. S., Wachterman, M., Weil, E., Stafford, R. S., Perrin, J. M., . . . & Iezzoni, L. I. (2005). The who, what, and why of risk adjustment: A technology on the cusp of adoption. *Journal of Health Politics, Policy and Law, 30*(3), 453–474.

Brown, R. S., Bergeron, J. W., Clement, D. G., Hill, J. W., & Retchin, S. M. (1993). *The Medicare Risk Program for HMOs: Final Summary Report on Findings from the Evaluation.* Technical report, Mathematica Policy Research.

Carter, G. M., Newhouse, J. P., & Relles, D. A. (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics, 9*(4), 411–428.

Center for Health Program Development & Management, & Actuarial Research Corporation. (2003). A Guide to Implementing a Health-Based Risk-Adjusted Payment System for Medicaid Managed Care Programs.

CMS. (2013, June). Key Milestones in CMS Programs. *Key Milestones in CMS Programs*. Retrieved from https://www.cms.gov/About-CMS/Agency-Information/History/index.html?redirect=/History/

CMS. (2014, December). *Medicare Shared Savings Program: Shared Savings and Losses and Assignment Methodology*. Retrieved from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Shared-Savings-Losses-Assignment-Spec-v2.pdf

Cumming, R., Knutson, D., Cameron, B., & Derrick, R. (2002, May). *A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations.* Technical report, Society of Actuaries.

Donabedian, A., Wheeler, J. R., & Wyszewianski, L. (1982). Quality, cost, and health: an integrative model. *Medical Care*, 20(10), 975–992.

Dunn, D. L., Rosenblatt, A., Taira, D. A., Latimer, E., Bertko, J., Stoiber, T., . . . & Busch, S. (1996, October). *A Comparative Analysis of Methods of Health Risk Assessment.* Technical report, Society of Actuaries.

Eggers, P. (1980). Risk differential between Medicare beneficiaries enrolled and not enrolled in an HMO. *Health Care Financing Review, 1*(3), 91–99.

Eggers, P. W., & Prihoda, R. (1982). *Pre-enrollment Reimbursement Patterns of Medicare Beneficiaries Enrolled in 'At-Risk' HMOs.* Office of Research, Office of Research and Demonstrations, Health Care Financing Administration.

Ellis, R. P., & Ash, A. (1995). Refinements to the diagnostic cost group (DCG) model. *Inquiry*, 32(4), 418–429.

Ellis, R. P., Ash, A., Consortium, H. P., & others. (1988). *Refining the Diagnostic Cost Group Model: A Proposed Modification to the AAPCC for HMO Reimbursement.* Brandeis Health Policy Research Consortium, Boston University.

Ellis, R. P., Pope, G. C., Iezzoni, L. I., Ayanian, J. Z., Bates, D. W., Burstin, H., & Ash, A. S. (1996). Diagnosis-Based Risk Adjustment for Medicare Capitation Payments. *Health Care Financing Review, 17*, 101–128.

Ellis, R. P., & others. (2000). Risk adjustment in competitive health plan markets. *Handbook of Health Economics, 1*, 755–845.

Feinstein, A. R. (1967). *Clinical Judgment.* Williams and Wilkins Co.

Fetter, R., Brand, D., & Gamache, D. (1991). *DRGs: Their Design and Development.* Health Administration Press.

Fetter, R. B. (1982). *The New ICD-9-CM Diagnosis Related Groups Classification Scheme: Sections I-IX: Final Report.* Health Systems Management Group, Yale University.

Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F., & Thompson, J. D. (1980). Case mix definition by diagnosis-related groups. *Medical Care*, 18(2), 1–53.

Fetter, R. B., Thompson, J. D., & Mills, R. E. (1976). A system for cost and reimbursement control in hospitals. *Yale Journal of Biology and Medicine, 49*(2), 123.

Fishman, P. A., Goodman, M. J., Hornbrook, M. C., Meenan, R. T., Bachman, D. J., & Rosetti, M. C. (2003). Risk adjustment using automated ambulatory pharmacy data: The RxRisk model. *Medical Care, 41*(1), 84–99.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press.

Geruso, M., & Layton, T. (2015). *Upcoding: Evidence from Medicare on Squishy Risk Adjustment*. National Bureau of Economic Research.

Gilmer, T., Kronick, R., Fishman, P., & Ganiats, T. G. (2001). The Medicaid Rx model: Pharmacy-based risk adjustment for public programs. *Medical Care, 39*(11), 1188–1202.

Goldfarb, M. G., & Coffey, R. M. (1992). Change in the Medicare case-mix index in the 1980s and the effect of the prospective payment system. *Health Services Research, 27*(3), 385.

Greenwald, L. M., Esposito, A., Ingber, M. J., & Levy, J. M. (1998). Risk adjustment for the Medicare program: Lessons learned from research and demonstrations. *Inquiry, 35*(2), 193–209. Retrieved from http://www.jstor.org/stable/29772755

Greenwald, L. M., Levy, J. M., & Ingber, M. J. (2000). Favorable selection in the Medicare+ Choice program: New evidence. *Health Care Financing Review, 21*(3), 127–134.

Hileman, G. (2015, December). A comparison of risk scoring recalibration methods. *Predictive Analytics and Futurism Section Newsletter,* Society of Actuaries.

Hileman, G., & Steele, S. (2016). *Accuracy of Claims-Based Risk Scoring Models.* Technical report.

Horn, S. D. (1981). Validity, reliability and implications of an index of inpatient severity of illness. *Medical Care*, 354–362.

Horn, S. D., & Schumacher, D. N. (1982). Comparing classification methods: Measurement of variations in charges, length of stay, and mortality. *Medical Care*, 489–500.

Hsu, J., Huang, J., Fung, V., Price, M., Brand, R., Hui, R., . . . & Newhouse, J. P. (2009). Distributing $800 billion: An early assessment of Medicare Part D risk adjustment. *Health Affairs, 28*(1), 215–225.

Iezzoni, L. I. (1997a). *Risk Adjustment for Measuring Healthcare Outcomes* (2nd ed.). Health Administration Press.

Iezzoni, L. I. (1997b). The risks of risk adjustment. *JAMA, 278*(19), 1600–1607.

Iezzoni, L. I. (2012). *Risk Adjustment for Measuring Healthcare Outcomes* (4th ed.). (L. I. Iezzoni, Ed.) Health Administration Press.

Iezzoni, L. I., & Moskowitz, M. A. (1986). Clinical overlap among medical diagnosis-related groups. *JAMA, 255*(7), 927–929.

Jencks, S. F., Williams, D. K., & Kay, T. L. (1988). Assessing hospital-associated deaths from discharge data: The role of length of stay and comorbidities. *JAMA, 260*(15), 2240–2246.

Kautter, J., Pope, G. C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., & Keenan, P. (2014a). The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. *Medicare & Medicaid Research Review, 4*(3).

Kautter, J., Pope, G. C., & Keenan, P. (2014b). Affordable Care Act risk adjustment: Overview, context, and challenges. *Medicare & Medicaid Research Review, 4*(3).

Kronick, R., T. Dreyfus, M. C., Lee, L., & Zhou, Z. (1998). Diagnostic risk adjustment for Medicaid: The disability payment system. *Managed Care: Advances in Financing, 17*(3), 3.

Kronick, R., Gilmer, T., Dreyfus, T., & Lee, L. (2000). Improving health-based payment for Medicaid beneficiaries: CDPS. *Health Care Financing Review, 21*(3), 29–64.

Kronick, R., Zhou, Z., & Dreyfus, T. (1995). Making risk adjustment work for everyone. *Inquiry*, 32(1), 41–55.

Lubitz, J. (1987). Health status adjustments for Medicare capitation. *Inquiry*, 24(4), 362–375.

Mehmud, S. (2013, July). *Nontraditional Variables in Healthcare Risk Adjustment.* Technical report, Society of Actuaries. Retrieved from https://www.soa.org/Research/Research-Projects/health/research-2013-nontrad-var-health-risk.aspx

Mehmud, S., & Yi, R. (2012, September). *Uncertainty in Risk Adjustment.* Technical report, Society of Actuaries. Retrieved from https://www.soa.org/research/research-projects/health/uncertainty-risk-adjustment.aspx

Mills, R., Fetter, R. B., Riedel, D. C., & Averili, R. (1976). AUTOGRP: An interactive computer system for the analysis of health care data. *Medical Care, 14*(7), 603–615.

NCHS. (1979). *International Classification of Diseases, 9th Revision, Clinical Modification, Volume 1 and 3.* NCHS.

Parkes, S. (2015, December). Producing actionable insights from predictive models built upon condensed electronic medical records. *Health Watch*.

Parkes, S., & Armstrong, B. (2015, July). Calibrating risk score models with partial credibility. *Forecasting and Futurism Section Newsletter,* Society of Actuaries.

Payne, S. (1987). Identifying and managing inappropriate hospital utilization. *Health Services Research, 22*, 709–769.

Pitches, D. W., Mohammed, M. A., & Lilford, R. J. (2007). What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Services Research, 7*(1), 91.

Pope, G. C., Bachofer, H., Pearlman, A., Kautter, J., Hunter, E., Miller, D., & Keenan, P. (2014). Risk transfer formula for individual and small group markets under the Affordable Care Act. *Medicare & Medicaid Research Review, 4*(3).

Pope, G. C., Ellis, R. P., Ash, A. S., Liu, C.-F., Ayanian, J. Z., Bates, D. W., . . . & Ingber, M. J. (2000). Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review, 21*, 93–118.

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., . . . & Robst, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review, 25*, 119–141.

Pope, G. C., Kautter, J., Ingber, M. J., Freeman, S., Sekar, R., & Newhart, C. (2011). Evaluation of the CMS-HCC risk adjustment model. RTI International and the Centers for Medicare & Medicaid Services (March).

Reid, R. J., Roos, N. P., MacWilliam, L., Frohlich, N., & Black, C. (2002). Assessing population health care need using a claims-based ACG morbidity measure: A validation analysis in the province of Manitoba. *Health Services Research, 37*(5), 1345–1364.

Relman, A. S. (1988). Assessment and accountability: The third revolution in medical care. *New England Journal of Medicine* (319), 1220–1222.

Robst, J., Levy, J. M., & Ingber, M. J. (2007). Diagnosis-based risk adjustment for Medicare prescription drug plan payments. *Health Care Financing Review, 28*(4), 15–30.

Romano, P. S. (2000). Should health plan quality measures be adjusted for case mix? *Medical Care, 38*(10), 977.

Romm, F. J., & Putnam, S. M. (1981). The validity of the medical record. *Medical Care*, 310–315.

Rosen, A., Reid, R., Broemeling, A., & Rakovski, C. (2003, May). Applying a risk-adjustment framework to primary Care: Can we improve on existing measures? *Annals of Family Medicine*, 1(1), 44-51.

Rosenberg, E. W. (1975). What kind of criteria. *Medical Care, 13*, 966–975.

Schone, E., & Brown, R. (2013, July). *Risk Adjustment: What Is the Current State of the Art and How Can It Be Improved?* Research Synthesis Report, Robert Wood Johnson Foundation.

Sonquist, J. A., & Morgan, J. N. (1964). *The Detection of Interaction Effects.* University of Michigan.

Starfield, B., Weiner, J., Mumford, L., & Steinwachs, D. (1991). Ambulatory care groups: A categorization of diagnoses for research and management. *Health Services Research, 26*(1), 53.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press.

U.S. National Library of Medicine. (2015). Drug Record: Nadolol.

Weiner, J. P., Starfield, B. H., Steinwachs, D. M., & Mumford, L. M. (1991). Development and application of a population-oriented measure of ambulatory care case-mix. *Medical Care, 29*(5), 452–472.

Weissman, J. S., Wachterman, M., & Blumenthal, D. (2005). When methods meet politics: How risk adjustment became part of Medicare managed care. *Journal of Health Politics, Policy and Law, 30*(3), 475–504.

Wennberg, J., & Gittelsohn, A. (1973). Small area variations in health care delivery: A population-based health information system can guide planning and regulatory decision-making. *Science, 182*(4117), 1102–1108.

WHO. (1975). *International Classification of Diseases, Vol. I.* World Health Organization.

Winkelman, R., & Damler, R. (2008). Risk adjustment in state medicaid programs. *Health Watch,* Society of Actuaries.

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

Young, W. W., Swinkola, R. B., & Zorn, D. M. (1982). The measurement of hospital case mix. *Medical Care*, 501–512.

Zimmer, J. G. (1974). Length of stay and hospital bed misutilization. *Medical Care, 12*, 453–462.