Supporting Healthcare Policy Initiatives through Modeling and Microsimulation Efforts: Issues of Data Capacity and Statistical Quality

Steven B. Cohen and Julie L. Hudson

Agency for Healthcare Research and Quality Working Paper No. 09002

August 2009

Suggested citation: Cohen SB, Hudson JL. Supporting Healthcare Policy Initiatives through Modeling and Microsimulation Efforts: Issues of Data Capacity and Statistical Quality . Agency for Healthcare Research and Quality Working Paper No. 09002, August 2009, http://gold.ahrq.gov.

Supporting Healthcare Policy Initiatives through Modeling and Microsimulation Efforts: Issues of Data Capacity and Statistical Quality

ABSTRACT

There is a growing demand for timely, high quality and precise estimates of health care parameters at the national and sub-national levels, and associated readily accessible data resources to inform health care policy and practice. Existing sentinel health care databases that provide nationally representative population based data on measures of health care access, cost, use, health insurance coverage, health status and health care quality, provide the necessary foundation to support descriptive and behavioral analyses of the U.S. health care system. Such studies help inform assessments of the availability and costs of private health insurance in the employment-related and non-group markets; the population enrolled in public health insurance coverage and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices.

To complement these assessments of the "current state" of health care, policymakers also depend on model-based estimates of the "future state" under alternative demographic, economic and technological assumptions, which are subject to greater levels of uncertainty traditionally associated with sampling and nonsampling error. Such modeling efforts directed to predicting the "future state" include economic models projecting health care expenditures and utilization, estimating the impact of changes in financing, coverage, and reimbursement policy, and determining who benefits and who bears the cost of a change in policy. Government and non-governmental entities rely upon these data to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare. Comparable standards of data quality and statistical integrity for these types of modeling and microsimulation efforts are needed to ensure policymakers have a sound understanding of the level of uncertainty associated with these model-based estimates. This presentation will focus on several of these ongoing health care modeling and microsimulation efforts to characterize sources of uncertainty in the resultant estimates and methodologies that can be employed to better quantify their error bounds.

Steven B. Cohen
Agency for Healthcare Research and Quality
Center for Financing, Access, and Cost Trends
540 Gaither Road
Rockville, MD 20850
Steven.Cohen@ahrq.hhs.gov

Julie Hudson
Agency for Healthcare Research and Quality
Center for Financing, Access, and Cost Trends
540 Gaither Road
Rockville, MD 20850
Julie.Hudson@ahrq.hhs.gov

1. Introduction

There is a growing demand for timely, high quality and precise estimates of health care parameters at the national and sub-national levels, and associated readily accessible data resources to inform health care policy and practice. Existing sentinel health care databases that provide nationally representative population based data on measures of health care access, cost, use, health insurance coverage, health status and health care quality, provide the necessary foundation to support descriptive and behavioral analyses of the U.S. health care system. Such studies help inform assessments of the availability and costs of private health insurance in the employment-related and non-group markets; the population enrolled in public health insurance coverage and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices.

To complement these assessments of the "current state" of health care, policymakers also depend on model-based estimates of the "future state" under alternative demographic, economic and technological assumptions, which are subject to greater levels of uncertainty traditionally associated with sampling and nonsampling error. Such modeling efforts directed to predicting the "future state" include economic models projecting health care expenditures and utilization, estimating the impact of changes in financing, coverage, and reimbursement policy, and determining who benefits and who bears the cost of a change in policy. Government and non-governmental entities rely upon these data to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare. Comparable standards of data quality and statistical integrity for these types of modeling and microsimulation efforts are needed to ensure policymakers have a sound understanding of the level of uncertainty associated with these model-based estimates (C. Citro and E. Hanushek, 1991). This paper will focus on the issues of data capacity and statistical quality to support modeling and microsimulation efforts. Particular attention will be given to the capacity of the Medical Expenditure Panel Survey to support these efforts and ongoing efforts to advance the utility of these

modeling efforts. Several examples of such modeling applications in concert with MEPS data to help inform health policy will also be presented.

2. Measurement of Trends in Health Care Cost, Coverage, Access and Use: MEPS Data Infrastructure

Health care expenditures represent nearly one-sixth of the United States gross domestic product, exhibit a rate of growth that exceeds other sectors of the economy, and constitute one of the largest components of the Federal and states' budgets. Although the rate of growth in health care costs slowed in the mid 1990s, it has recently begun to rise again, fueled primarily by increasing costs for hospital care and prescription medications. As a result, the question of how to design a system that encourages the provision of high quality care as efficiently as possible remains an issue of continuing concern to both private and public payers. In a similar vein, an evaluation of the current health care system requires an understanding of the patterns and trends in the use of health care services and their associated costs and sources of payment. To effectively address these issues, researchers and policymakers need accurate nationally representative data to better permit an understanding of how individual characteristics, behavioral factors, financial incentives, and institutional arrangements affect health care utilization and expenditures in a rapidly changing health care market.

The growing demand for accurate and reliable information on the population's health care utilization, expenditures, insurance coverage, sources of payment and access to care served as the catalyst to initiate the implementation of the family of national medical expenditure surveys sponsored by the Agency for Healthcare Research and Quality (AHRQ) and its predecessor agencies. AHRQ's Medical Expenditure Panel Survey (MEPS) collects detailed information regarding the use and payment for health care services from a nationally representative sample of Americans. It is also co-sponsored by the National Center for Health Statistics, CDC. Westat is the data collection organization.

The MEPS research program, broadly defined to encompass data collection, data development, research and the translation of research into practice, is directly tied to the strategic goal of identifying strategies to improve access, foster appropriate use and

reduce unnecessary expenditures. Few other surveys provide the foundation for estimating the impact of changes on different economic groups or special populations of interest, such as the poor, elderly, veterans, the uninsured, or racial/ethnic groups. The public sector relies upon the MEPS research findings to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare. In the private sector, these data are also used to develop economic projections.

Over the past several years, the data and associated research findings have quickly become a linchpin for the nation's economic models and their projections of health care expenditures and utilization. This combination of breadth and depth of the data enables public and private sector analysts to develop economic models designed to produce national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy. Since 1977, AHRQ's expenditure surveys have been an important and unique resource for public and private sector decision makers. The survey is unique in the level of detail of information obtained on the health care services used by Americans at the household level and their associated expenditures (for families and individuals); the cost, scope, and breadth of private health insurance coverage held by and available to the U.S. population; and the specific services purchased through out-of-pocket and/or third-party payments. The Medical Expenditure Panel Survey (MEPS) is the most recent expenditure survey effort, initiated in 1996, and designed as a continuous on-going survey to permit annual estimates of health care utilization, expenditures, insurance coverage and sources of payment for the U.S. civilian noninstitutionalized population. The MEPS data support a wealth of basic descriptive and behavioral analyses of the U.S. health care system. These include studies of the population’s access to, use of, and expenditures and sources of payment for health care; the availability and costs of private health insurance in the employment-related and non-group markets; the population enrolled in public health insurance coverage and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices.

The MEPS consists of a family of three interrelated surveys: the Household Component (HC), the Medical Provider Component (MPC), and the Insurance Component (IC). The MEPS Household Component was designed to provide annual national estimates of the health care use, medical expenditures, sources of payment and insurance coverage for the U.S. civilian non-institutionalized population. In addition to collecting data to yield annual estimates for a variety of measures related to health care use and expenditures, MEPS also provides estimates of measures related to health status, demographic characteristics, employment and access to health care. Estimates can be provided for individuals, families and population subgroups of interest. The MEPS HC consists of an overlapping panel design in which any given sample panel is interviewed a total of 5 times in person over 30 months to yield annual health insurance coverage, use and expenditure data for two calendar years. These rounds of interviewing are spaced about 5 to 6 months apart. The interview is administered through a computer assisted personal interview mode of data collection, and takes place with a family respondent who reports for him/herself and for other family members. The most current MEPS annual survey consists of approximately 14,000 families and 35,000 individuals, and reflects an oversample of the following policy relevant population subgroups: Hispanics, blacks, Asians and low income households. Data from two panels are combined to produce estimates for each calendar year. The data collected in this ongoing longitudinal study also permit studies of the determinants of the use of services and expenditures, and changes in the provision of health care in relation to social and demographic factors such as employment or income; the health status and satisfaction with health care of individuals and families; and the health needs of specific population groups such as the elderly and children. The MEPS Medical Provider Component (MEPS-MPC) is a supplement to the household component designed to provide additional information on charges and sources and amounts of payment received by providers for care delivered to participants in the household survey.

The MEPS Insurance Component (IC) is a survey of private business establishments and governments designed to obtain information on health insurance availability, coverage, and cost derived from employers in the U.S. The sample for this survey is selected from the Census Bureau's Business Register for private employers and

Census of Governments for public employers. The IC is an annual survey designed to provide both nationally and state representative data on the types of health insurance plans offered by employers, enrollment in plans by employees, the amounts paid by both employers and employees for those plans, and the characteristics of the employers.[i] The data are collected by the Census Bureau and are protected under the confidentiality provisions of Title 13 (the Bureau's authorizing legislation). As a result, IC data are disseminated publicly only through summary data tables posted on the AHRQ website.[ii]

Health care policymakers require accurate estimates of the size and composition of the insured and uninsured populations, as well as information on how demographic characteristics, economic factors and health status affect health plan eligibility and decisions to enroll in health insurance plans. Furthermore, efforts to address inequities in the availability of private health insurance and to control health insurance premiums and medical care costs must necessarily focus on the employment-related health insurance market. Historically, the analyses of data from the MEPS family of surveys has figured prominently in this arena. As is evidenced in the recent Institute of Medicine (IOM) Report on "Health Insurance is a Family Matter", the report notes that "the most comprehensive data on who uses what health care service and how much is paid for those services comes from the Medical Expenditure Panel Survey" (IOM, 2002). MEPS related analyses are prominently used to inform components of this IOM report focused on issues of insurance coverage and cost.

MEPS derived estimates of the health insurance status of the U.S. civilian non-institutionalized population are critical to policymakers and others concerned with access to medical care and the cost and quality of that care. Health insurance helps people get timely access to medical care and protects them against the risk of expensive and unanticipated medical events. When estimating the size of the uninsured population, it is critical to consider the distinction between those uninsured for short periods of time and those who are long term uninsured across several years in duration. Compared to people with healthcare coverage, uninsured people are less likely to visit a doctor, have a usual source of medical care, receive preventive services, or have a recommended test or prescription filled Consequently, individuals that experience extended periods of being uninsured are particularly at risk for restrictions in access to care and exposure to serious

illness and significant financial jeopardy. Since many individuals undergo transitions in the acquisition and loss of health insurance coverage over time, an important consideration is the length of duration of spells of un-insurance and the capacity of this lack of coverage to lead to less efficient use of health care services and facilities. In this regard, MEPS research efforts have demonstrated that individuals who experience short spells of being uninsured differ significantly from those who have been uninsured for more than a year on several dimensions which include access to employer sponsored coverage, their attitudes and preferences regarding the need for coverage and their sensitivity to the cost of acquiring coverage. In addition to providing cross-sectional estimates of health insurance coverage each year, the MEPS has the added analytical capacity to identify individuals with gaps in coverage over time as well as the duration of the spells of being uninsured for up to four years in duration.

In addition to measuring actual out-of-pocket financial burdens for health care, MEPS provides the only nationally representative data that can be used to measure the extent of underinsurance in the U.S. Underinsurance is defined as being at risk of spending more than a certain amount of family income on out of pocket expenses in the event of a catastrophic medical illness. Estimates of the underinsured require linked information on families health insurance benefits, family income, and risk of experiencing catastrophic medical events that are found in the MEPS.

With health care absorbing increasing amounts of the nation's resources, the question of how to design a system that encourages the provision of high quality care as efficiently as possible remains an issue of continuing concern to both private and public payers. To effectively address this issue, researchers and policymakers have benefited from MEPS research findings to better understand how individual characteristics, behavioral factors, financial incentives, and institutional arrangements affect health care expenditures in a rapidly changing health care market. Research findings for the MEPS have also served to provide health care decision makers with a better understanding of the highly concentrated nature of health care expenditures and the persistence of these high expenditures over time. MEPS studies that examine the persistence of high levels of expenditures over time have been essential to help discern the factors most likely to drive health care spending and the characteristics of the individuals who incur them.

Furthermore, additional attention and prioritization has been given to data collection procedures, predictive modeling and estimation strategies that help improve the precision and quality of the survey estimates that characterize this policy relevant population subgroup of individuals with high levels of medical expenditures. Research findings from MEPS also provide clear evidence of the utility and appropriateness of probabilistic models as prediction tools for identifying individuals likely to incur high levels of medical expenditures in future years. To the extent that this policy relevant subset of the population is amenable to successful prediction through the application of well developed models, the methodology continues to find several venues for application. Prominent examples of applications ripe for implementation include adoption of oversampling strategies for national health care surveys, and the identification of individuals whose health status improvements through disease management programs could most significantly impact on potential reductions in overall future year health care expenditures.

Given the growing attention being given to achieving a better understanding of the impact of rising prescribed medicine costs on health and the consumption of health services, it is also important to note the utility of the MEPS to inform studies examining the association between the use of newer medicines and morbidity, mortality, and health spending. Using this data resource, researchers have been able to determine the direction of the association between the use of newer drugs and all other types of nondrug medical spending. Attention has also focused on studies that identify inappropriate medication use, which is a major patient safety concern and has significant consequences with respect to health care costs. With its wealth of data on health conditions, prescribed medication utilization and expenditures and associated therapeutic drug classifications, the MEPS data have also been helpful to researchers attempting to identify potentially inappropriate medication use in the community.

2.1 How the research shaped or influenced decisions, policy formulation or public discourse: examples of impact on health care policy and practice

Since its inception, MEPS has been used in several hundred scientific publications, and many more unpublished reports. It has served a core data resource for

estimating the impact of changes on different economic groups or special populations of interest, such as the poor, elderly, veterans, the uninsured, or racial/ethnic groups. Government and non-governmental entities rely upon these data to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare. In the private sector (e.g., RAND, Heritage Foundation, Lewin-VHI, and the Urban Institute), these data are used by many private businesses, foundations and academic institutions to develop economic projections. These data represent a major resource for the health services research community at large. Since 2000, data on premium costs from the MEPS Insurance Component have been used by the Bureau of Economic Analysis to produce estimates of the GDP for the nation. The data are also used to inform the national health care cost estimates in the National Health Expenditure Accounts and to assess time trends in the provision of employer health benefits by States. Additional examples of how the research has shaped or influenced decisions and policy formulation follow:

- The MEPS research findings have been used extensively by the Congressional Budget Office, Department of Treasury, Joint Taxation Committee and Department of Labor to inform Congressional inquires related to health care expenditures, insurance coverage and sources of payment and to analyze potential tax and other implications of Federal Health Insurance Policies.
- MEPS research findings on health care quality, access and health insurance coverage have been used extensively in the Department's two annual reports to Congress, the National Healthcare Disparities Report and the National Healthcare Quality Report.
- The MEPS has been used in Congressional testimony on the impact of health insurance coverage rate increases on small businesses.
- The MEPS data have informed studies of the value of health insurance in private markets and the effect of consumer payment on health care,
- The MEPS-IC has been used by a number of States in evaluating their own private insurance issues including eligibility and enrollment by the State of Connecticut and by the Maryland Health Care Commission; and community rating by the State of New York. As part of the Robert Wood Johnson Foundation's State Coverage Initiative, MEPS data was cited in 69 reports, representing 27 states.
- The MEPS has been used extensively by the Government Accountability Office to determine trends in Employee Compensation, with a major focus on the percentage of employees at establishments that offer health insurance, the percentage of eligible employees who enroll in the health insurance plans, the average annual premium for employer-provided health insurance for single

workers, and the employees' share of these premiums.

- MEPS has been used in DHHS Reports to Congress on expenditures by sources of payment for individuals afflicted by conditions that include acute respiratory distress syndrome, arthritis, cancer, chronic obstructive pulmonary disease, depression, diabetes, and heart disease.
- MEPS is used to develop estimates provided in the *Consumers Checkbook Guide to Health Plans*, of expected out of pocket costs (premiums, deductibles and copays) for Federal employees and retirees for their health care. The *Checkbook* is an annual publication that provides comparative information on the health insurance choices offered to Federal workers and retirees.
- MEPS has been extensively used to examine the pharmacological treatment of many conditions including depression (in both adults and children), back pain, ADHD, obesity, hypertension and cardiovascular diseases.
- MEPS has been used by CDC and others to evaluate the cost of common conditions including arthritis, injuries, diabetes, obesity and cancer.
- MEPS has been used to examine quality of care, including the receipt of preventive care and barriers to that receipt. MEPS data has been used by private sector insurance firms to estimate the potential return on investment to firms for providing bariatric surgery benefits to their enrollees.
- After the enactment of SCHIP, the MEPS data were used to analyze the impact of insuring uninsured children. Updates of the estimates of the number of eligible uninsured children helped inform policy formulation on outreach and program cost.

Longitudinal Capacity: Research efforts build on the analytical strengths of the MEPS to support longitudinal analyses and takes advantage of its integrated survey design linked to the National Health Interview Survey to expand the time period and analytical profiles of the sample respondents to these integrated surveys. With the MEPS longitudinal design, analysts have assessed the persistence of high health care expenditures by examining whether individuals in high expenditure percentiles of the health care expenditure distribution in a given year remain in upper percentiles in the following year or shift to another higher or lower percentile. The overlapping panel design of the MEPS has also been used to assess the impact of survey attrition on the resultant survey estimates by comparing the national health care estimates produced by the first year of a sample panel (with a higher response rate) in contrast to the estimates derived from the second year of a MEPS sample panel covering the same time period. In addition, with the linkage of MEPS and NHIS files, longitudinal analyses of transitions in health insurance coverage and health status characteristics have be examined over a 3-

year period. All the survey estimates and analyses conducted with the MEPS adjust for survey design complexities and include adjustments for survey nonresponse and poststratification. The survey and resultant analyses have markedly benefited by the conduct of ongoing statistical and methodological research initiatives to improve the accuracy, precision, efficiency, timeliness and overall data quality and analytical capacity of the survey.

2.2 Analytical enhancements achieved through linkage of surveys to other sources of data

The analytical capacity of health surveys such as the MEPS are dramatically enhanced through the linkage to existing secondary data sources at higher levels of aggregation (both geographic and organizational) as well as through direct matches to additional health and socio-economic measures acquired for the same set of sample units from other sources of survey specific or administrative data (Cohen et al., 2005). One of the more pervasive uses of existing administrative data bases is to serve as a sampling frame to facilitate a cost efficient identification of an eligible survey population for purposes of sample selection, such as the consideration of the Medicare administrative records to serve as a sampling frame for a survey of Medicare beneficiaries. Health surveys that are so linked to administrative records from their inception benefit by this capacity for data supplementation that permits enhanced and more extensive analyses that are beyond the more constrained scope of the core health survey. Establishing similar connections to existing data sources that will substantially enhance a survey's capacity to address specific research questions is often more difficult to establish after a survey has been administered. This is primarily a consequence of confidentiality restrictions that require respondent permission to link patient records to administrative data sources, in addition to problems with the availability of the necessary identifiers from the survey respondents.

The large majority of the nationally representative population-based health surveys sponsored by the Department of Health and Human Services have benefited by a capacity to link the survey data to county level data on health service resources and health manpower statistics available on the Area Resources File (ARF). More

specifically, the ARF is a county-specific health resources information system containing information on health facilities, health professions, measures of resource scarcity, health status, economic activity, health training programs, and socio-economic and environmental characteristics. Geographic codes and descriptors are provided to enable linkage to health surveys to expand analyses conducted by planners, policymakers, researchers, and other professionals examining the nation's health care delivery system and factors that may impact health status and health care in the U.S. Comparable enhancements to health surveys for supplementation of economic indicators are achievable through linkage of survey data to the socio-economic indicators made available by the Bureau of the Census through the County and City Data Book and public use files from the decennial Census.

The quality and data content of household specific health surveys are often enhanced through the conduct of follow back surveys to medical providers and facilities that have provided care to household respondents. In terms of data quality, household reported medical conditions can be evaluated for accuracy relative to provider specific records on medical conditions for the same patient and specific health events. With respect to health care expenditures collected from household respondents for their reported health care events, available linked medical provider level data is a more accurate source of information. The availability of such supplemental data on use and expenditures allows for the conduct of methodological studies to evaluate the accuracy of household reported data and informs adjustment strategies to household data in the absence of provider specific data to reduce bias attributable to response error.

3. Data Capacity and Statistical Quality Considerations in Support of Modeling Efforts

The MEPS also includes the continued development and updating of the MEDical expenditure microSIMulation (MEDSIM) project. The MEDSIM project consists of several modular simulation programs and complex databases that enhance the capabilities of the MEPS to analyze, track, and project changes in the health care system. These modular units are designed with flexibility to support a variety of analyses regarding economic impacts of current health care policies as well as possible future

modifications. These research tools have been designed to answer a broad range of "what if" questions regarding the impacts of possible changes in the financing and delivery of health care services.

When undertaking such modeling efforts to address these "what if" scenarios, it is important to insure that the principal policy-maker(s) receive a clear understanding of the resultant study findings and applicability, underlying model assumptions, sources of uncertainty and their relative impacts on stability of the estimates. To help achieve this objective, it is imperative that the following set of attributes of the modeling enterprise are appropriately addressed:

- Selection of Host Analytic Database/Data Capacity
- Model Specifications
- Analytical and Statistical Oversight
- Methodology
- Precision
- Nonsampling Errors
- Evaluation of Performance
- Reproducibility of results
- Limitations
- Transparency

Selection of Host Analytic Database/ Data Capacity: A major consideration that needs to be addressed at the outset is the identification of the most appropriate data source or set of data resources to inform the primary analyses. Issues of content, national and sub-national representativeness, sample size and data quality, timeliness, and accessibility are sentinel to the decision regarding the choice of "host" data set. Given the unique features that characterize the family of nationally representative health care surveys that are currently in the field, the choice is often straight forward, frequently driven by analytic content and scale. Alternatively, a more complex set of trade-offs are considered when the modeling effort must address several policy questions simultaneously.

The following example is provided to help illustrate the trade-offs encountered in the selection of a host dataset, when there are competing objectives inherent in the

primary analyses. Suppose the question at hand was to evaluate the impact of an increase in the percent of the population with incomes between 200%-400% of the poverty level with health insurance coverage on medical expenditures, by state. Use of the MEPS would be the appropriate choice for national level impacts, but would not have the power to directly provide estimates for the smaller states based on population. Small area estimation techniques would need to be applied to facilitate national estimates, resulting in increases in mean square error estimates as a consequence of additional sources of bias in estimates. Alternatively, the CPS would have information on both coverage and income, at the state level, but would need to acquire the medical expenditure data from the MEPS through many-to-one statistical matching or "cold deck" imputation strategies. Mean square error estimates derived from this strategy would be increased as a consequence of the impact of imputation on variance estimation as well as additional sources of bias in estimates attributable to the matching/imputation.

Model Specification and Capacity: An a priori decision is necessary at the outset regarding the scope and capacity of the modeling effort. A model that is developed in response to addressing a distinct set of highly related health care policy questions will achieve gains in specificity at the expense of flexibility and utility for addressing a more expansive set of policy initiatives. Alternatively, a more broadly configured modeling effort will primarily be optimized to address a more general set of integrated health care policy initiatives, but will need more frequent and intensive modifications to address specific policy changes. The process of model development must follow conventional statistical and econometric techniques for model selection, choice of model inputs, determination of significant factors, assessments of model performance and should be subject to model validation.

For modeling efforts that need to address the impact of a policy change in the future, decisions must be made at the outset regarding whether a static or dynamic approach to aging the data will be taken. In addition, decisions regarding the incorporation of modeling the impact of behavioral responses in response to the policy

are essential to defining model capacity and scope and impact on level of complexity, cost and time for development.

Analytical and Statistical Oversight: The modeling effort should be based on sound statistical theory, rigor and practice. The products of the modeling effort should be subject to a rigorous statistical and substantive review to insure statistical standards developed by the statistical agencies responsible for the core data sources to guide external users in the appropriate uses of the data and restrictions have been satisfied.

A well informed modeling effort imposes an interdependence between the modelers, the statisticians and methodologists on the project, statisticians and researchers associated with the host dataset, the data processing staff and the end users, who are primarily health researchers, policymakers and the public. When all the essential contributors to the modeling effort work in concert, the overall quality and utility that are achieved is enhanced. .

Methodology: A report that describes the underlying approach taken to determine the final model specification, and summarizes the results of statistical tests employed to determine model fit and error is essential. The estimation strategies employed to age the data to reflect the demographic configuration of the target population into the future, under either static or dynamic approaches should also be clearly specified to facilitate understanding and replication. In addition, when the model is developed to assess the impact of behavioral responses in response to a policy change, the underlying economic theory and empirical evidence that supports the approach taken to address this dimension of the simulation needs to be documented in addition to the explicit model specifications that were adopted.

Precision: The sample design of the host datasets used for modeling efforts is often characterized by a complex multi-stage area probability design that includes disproportionate sampling of specified policy relevant population groups. As a consequence of departures from simple random sampling assumptions, the variances of

the model based estimates derived from the host survey must take into account the complex survey design. Consequently, the error estimates that characterize these model-based estimates should reflect the component of variance attributable to the underlying sample and survey design. Furthermore, imputation is often used to correct for item nonresponse associated with measures that are essential for model specification. The overall variance of model based estimates should also incorporate an additional component of variance attributable to imputation. Additional sources of variation associated with population projections used in re-weighting to "age" the data need to also be accounted for. When all these components are accounted for in deriving estimates of the precision of the model-based estimates, this should be considered as a lower bound when attempting to characterize the overall mean square error to convey a measure of uncertainty in the results.

Nonsampling Errors: In addition to sampling errors, several components of error impact on the quality of the estimates derived from the host dataset used for modeling purposes. They include errors associated with model specification, nonresponse, measurement, and population projections, in addition to errors in the coverage of the target population. The measurement and reporting of these sources of error, and the derivation of estimates of their scale and impact on bias in estimates is important for everyone who uses the host dataset(s) and the modeling tool. The resultant estimate of bias is another component that needs to be incorporated in the mean square error associated with the model's findings.

Evaluation: Once model specification is completed, the model should undergo an evaluation phase to better assess accuracy in prediction. An examination of the performance of a model with the same data used to develop the best predictive model specification risks contamination of the evaluation. A model validation assessment is best performed through application to an independent representative sample that characterizes the same target population. This condition is often satisfied through development of the prediction model using data from a representative half-sample, and then applying the model to the other representative half-sample to assess model performance. Alternatively, when focused on modeling efforts that require future projections, this condition could be

satisfied through development of the prediction model using data from a longitudinal host data set. The evaluation would then focus on the application to an independent future longitudinal panel obtained from the same dataset, where a direct estimate of the criterion measure(s) was available, to assess model performance. The robustness of model based estimates should be further assessed by examining the alignment in estimates when using alternative competitor models and host data sets. Follow-up assessments over time should also be conducted to examine the accuracy of the model-based estimates with actual results to help refine model specifications and clarify which factors were not appropriately accounted for. Independent evaluations conducted by experts distinct from the model developers help bring an additional level of scrutiny and quality assurance to the process.

Reproducibility of results: Sufficient documentation should be provided to facilitate the reproducibility of model-based results by independent application of the modeling program(s). This capacity will permit the conduct of replication studies, additional data analysis, and sensitivity testing to help reconcile observed differences resultant estimates obtained from alternative approaches. The dissemination of the model source code, the application software and the underlying host dataset that has been "aged" and modified to satisfy model inputs would serve as an effective and efficient strategy to achieve this aim. These dissemination efforts will often be constrained, in order to insure the confidentiality provisions of the host survey(s).

Limitations: Limitations in model capacity and measures of the level of uncertainty associated with resultant model-based estimates should be clearly stated to facilitate a good understanding of the risks inherent in policy decisions based on model outputs. To the extent possible, metrics should be provided to provide the user with an indication of the amount of overall uncertainty in the estimates and the respective contributions to error attributable to sampling variability, macroeconomic forecasts, imputation routines, and modeling approaches.

Transparency: Attention should also be given to the provision of good documentation of all the steps involved in model development, specification, statistical testing,

performance, evaluation, application and interpretation of results to help insure the credibility, integrity and accessibility of the modeling tool. Within this framework, a full description of the data collection and processing methods, including data editing, imputation methods, and weighting methods should be conveyed to model users and consumers. Clarifications of survey data collection phenomena that contribute to potential sources of error should be included.

Good documentation is required to permit new users to gain a thorough understanding of the model, a capacity to use and/or modify it in a cost-effective manner and an ability to further evaluate the impact of model components and their effects on the quality of the model output. To the extent possible, documentation should also include the program code and audit trails that track the effects of modifications in model components.

4. Specific Examples of Health Care Modeling Efforts Using MEPS

4.1 Projected Medical Expenditures by Type of Service and Source of Payment for Age, Race and Sex through 2016: Data from the National Health Expenditure Accounts (NHEA)-Aligned Medical Expenditure Panel Survey (MEPS)

Projected future health care expenditures derived from MEPS are calculated in a two step process (J. Hudson, 2008)[1]. First, core data from the 2002 NHEA-aligned MEPS file are projected to each end year through 2016 by adjusting MEPS person weights using Census data[2] on population growth characteristics over time. Then, the re-weighted NHEA-aligned MEPS expenditures are calibrated annually by type of service and source of payment categories so that growth in the re-weighted NHEA-aligned MEPS

[1] For more details on NHEA-aligned MEPS, see: Sing et. al., 2006. "Reconciling Medical Expenditure Estimates from the MEPS and NHEA, 2002," *Health Care Financing Review* and Selden and Sing, 2008 "Aligning the Medical Expenditure Panel Survey to Aggregate U.S Benchmarks," AHRQ unpublished manuscript (available upon request from jhudson@ahrq.gov).

[2] U.S. Interim Projections by Age, Sex, Race, and Hispanic Origin (based on the 2000 Census) is issued by the Population Projections Branch at the Census Bureau (usproj2000-2005.csv). This file can be downloaded from the Census Bureau website from the section entitled "Detail file" on the following page: http://www.census.gov/ipc/www/usinterimproj/ (accessed December 1, 2007).

expenditures matches growth in the projected NHEA[3, 4]. The projection process yields a set of annual data files and summary statistics available for public use for all future years in which the MEPS-HC Full Year Consolidated File is not available. These files[5], contain person level data with (i) individual level medical expenditures classified by type of service and source of payment categories, (ii) a person weight that enables national-level estimates, (iii) individual level demographic characteristics used in the re-weighting procedure and iv) insurance and poverty status variables, not controlled for in the aging process[6].

Each person record contains data on the expenditures made within seven different types of service categories plus a total expenditure category. Total expenditures are calculated by adding expenditures for all seven categories[7]. The categories and their two to three character abbreviations are as follows:

1. Hospital: HOS
2. Physician: PHY
3. Other Medical: OTH
4. Home Health: HHC
5. Other Provider: OBO
6. Dental: DVT
7. RX: RX

[3] CMS's 2005 figures from file NHE65-16.CSV are used for NHEA expenditures. In NHE65-15.CSV, data for 1965-2005 represent historical expenditures, while data for 2006-2016 are projected expenditures. This data file is available to the public on the CMS web site by using the link entitled "The NHE Historical and Projections 1965-2016" on the following page: http://www.cms.hhs.gov/NationalHealthExpendData/03_NationalHealthAccountsProjected.asp#TopOfPage (accessed December 1, 2007).

[4] In order to calibrate NHEA-aligned MEPS to the NHEA, nursing home and assisted living expenditures by type of service and source of payment first had to be added to the NHEA-aligned MEPS expenditures because the survey does not collect these expenditures. There will be more detail on this adjustment in section 4 of this documentation.

[5] Where yy is a two-digit year code. For example, the projected data file for 2015 has the name PMEP15.DAT.

[6] Researchers should use these variables with caution. The aging process for the NHEA-aligned MEPS does not control for changes in these measures, so any estimates with the projected data cross-classified by the poverty status or health insurance status variables are based on the assumption that 2002 poverty and health insurance status conditions remain unchanged in each future year. Any changes in poverty rates or health insurance status of individuals observed in the projected NHEA-aligned MEPS are an indirect result of the changing age, race and sex composition of the population that was directly controlled for in aging the data. These variables are included in the projected data sets purely for the convenience of end-users.

[7] TOT = HOS + PHY +OTH + HHC + OBO + DVT + RX.

8. TOTAL: TOT[8]

The expenditures in each category are divided into nine sources of payment, plus a total. Total expenditures (EXP) equal the sum of the nine source of payment expenditure variables[9]. Those categories and their two to three character abbreviations are as follows:

1. Out of Pocket: SLF
2. Private Health Insurance: PHI
3. Medicare: MCR
4. Medicaid: MCD
5. Tricare: TRI
6. VA: VA
7. Workers Comp: WC
8. Other Public: OTP
9. Other: OSR
10. TOTAL: EXP[10]

There are a total of 80 expenditure variables for all the type of service variables crossed with all the source of payment variables above.

Population Aging: Each person record includes a person-level sample weight needed to produce estimates for the civilian, non-institutionalized population of the US in the designated year. The weight on each file is derived from the 2002 NHEA-aligned MEPS weight. It is adjusted to reflect changes in the non-institutionalized civilian population between 2002 and the designated end year. Weights are adjusted to reflect population growth by age, race/ethnicity and sex based on the U.S. Interim Projections by Age, Sex, Race, and Hispanic Origin issued by the Population Projections Branch at the Census Bureau. Individuals in the 2002 NHEA-aligned MEPS sample are placed into cells based on age, race/ethnicity and sex categories to match definitions in the Census Bureau's interim population projections file. Population growth rates are calculated for each age, race/ethnicity, sex cell from 2002 to the designated end year using the Census interim projected population totals by age, race/ethnicity and sex. These growth rates are then applied to the person-level weights in the 2002 NHEA-aligned MEPS file by age, race/ethnicity and sex.

[8] Note that Total Expenditures (TOT) by type of service equals Total Expenditures (EXP) by source of Payment.
[9] EXP = SLF + PHI + MCR + MCD + TRI + VA + WC + OTP + OSR.

Calibration to National Health Accounts: The last step of the aging process is a calibration of each re-weighted NHEA-aligned MEPS file to the National Health Expenditure Accounts for the same year by type of service and source of payment. This process ensures that growth rates in NHEA-aligned MEPS expenditures by type of service and source of payment between 1996 and a designated end year will be in accordance with growth rates predicted by the NHEA.

In order to calibrate NHEA-aligned MEPS to the NHEA, it is first necessary to obtain expenditures for nursing home and assisted living for 2002, because MEPS does not include these expenditures in the survey. These totals were obtained from Actuarial Research Corporation (ARC)[10] by type of service and source of payment for 2002. For years beyond 2002, the ARC totals are aged by the growth rates projected by CMS in the National Health Expenditure Accounts for nursing home expenditures by type of service and source of payment.

The calibration procedure is as follows for each designated end year: (i) Expenditures from the re-weighted NHEA-aligned MEPS file are totaled by the seven type of service categories and the nine source of payment categories for a total of 63 TOS-SOP cells for a given end year. (ii) Nursing home and assisted living expenditures are added to the 2002 NHEA-aligned MEPS expenditures by type of service and source of payment. (iii) "Aged" nursing home and assisted living expenditures for the designated end year are added to the re-weighted NHEA-aligned MEPS file by type of service and source of payment. (iv) Growth rates from 2002 to the designated end year are calculated for each of the TOS-SOP categories for the NHEA-aligned MEPS re-weighted data. (v) Growth rates are calculated for the corresponding TOS-SOP categories for the NHEA using historical and projected data based on 2005 estimates from CMS. (vi) Using the

[10] Nursing home and assisted living expenditures by type of service and source of payment: Unpublished data made available to AHRQ and CMS by ARC for the completion of Sing et. al. 2006. Adjustments to the original ARC estimates were made by a team of researchers at AHRQ, CMS and ARC to best match the needs of reconciling the 2002 MEPS and 2002 NHEA expenditures.

NHEA and NHEA-aligned MEPS growth rates, a calibration factor[11] is created for each TOS-SOP category for the designated end year. (vii) Expenditures in the re-weighted NHEA-aligned MEPS file are multiplied by TOS-SOP specific calibration factors.

Additional Data Details and Caveats: It is important to emphasize that there are several sources of error associated with the interim MEPS projected expenditures and with any estimates produced using these data. These data are surrounded by several sources of uncertainty. Sources of error include sampling error in the MEPS and NHIS files, variation resulting from the data collection and projection techniques used by Census Bureau and CMS in creation of the Census and NHEA files, uncertainty in ARC estimates of the nursing home and assisted living population, variation resulting from the alignment of the 2002 MEPS and 2002 NHEA files and, finally, variation from the procedures used to age the MEPS population and calibrate aged expenditures to NHEA growth.

The interim nature of the data reflects the use of interim population projections based on the 2000 U.S. Census[12]. The interim Census file projects annual population totals by age, sex, race and Hispanic origin for only a subset of racial groups[13] and does not contain the components of change for fertility and mortality historically used in projected MEPS expenditures. Therefore, the interim projections associated with this data release will be replaced when final Census population projections become available[14].

[11] The calibration factor is calculated as follows: (growth rate for NHEA expenditures/growth rate for population aged NHEA-aligned MEPS expenditures). Note there is a different calibration factor for each designated end year and for each TOS-SOP category. See Table 2 in the appendix for the corresponding NHEA and NHEA-aligned MEPS expenditures used for the calibration.
[12] U.S. Interim Projections by Age, Sex, Race, and Hispanic Origin issued by the Population Projections Branch at the Census Bureau (usproj2000-2005.csv). This file can be downloaded from the Census Bureau website from the section entitled "Detail file" on the following page: http://www.census.gov/ipc/www/usinterimproj/ (accessed December 1, 2007).
[13] The set of racial and ethnic groups available in the interim Census file are not compliant with requirements set by the Office of Management and Budget with respect to mixed races and does not allow researchers to distinguish non-white race population totals by Hispanic Origin.
[14] Final Census population projections are expected to be released to the public in 2008.

Researchers should note that, in addition to the interim nature of these data, the current release of projected MEPS expenditures differs from those in previous years. Projected expenditures from the 1996 MEPS were calculated directly from the 1996 MEPS-HC Full Year Consolidated Public Use File (HC-012). By using the 2002 NHEA-aligned MEPS instead of the 2002 MEPS Full Year Consolidated File (HC-070), the current release has been improved and takes into account differences found between the 2002 MEPS and NHEA totals.

4.2 Use of Probabilistic Models to Oversample the Long Term Uninsured

Given the risk of exposure to high out of pocket medical expenditures faced by the long term uninsured and associated economic and health related consequences, this population subgroup is of particular relevance to health policy considerations. Consequently, a prediction model that can accurately identify the long term uninsured is an important analytical tool. These models have particular relevance as statistical tools to facilitate efficient sampling strategies that permit the selection of an over-sample of individuals likely to be uninsured for long periods in the future (S. Cohen and W. Yu, 2008; S. Cohen et al. 2006, 2003). This example provides a summary of the development of prediction models to identify the long term uninsured adults under age 65 and includes an evaluation of its potential utility as an oversampling strategy for use in the Medical Expenditure Panel Survey (MEPS).

Compared to people with healthcare coverage, uninsured people are less likely to visit a doctor, have a usual source of medical care, receive preventive services, or have a recommended test or prescription filled (Selden and Hudson, 2006; Taylor et al. 2001; Weinick et al. 1996). Consequently, individuals that experience extended periods of being uninsured are particularly at risk for restrictions in access to care and exposure to serious illness and significant financial jeopardy. Since many individuals undergo transitions in the acquisition and loss of health insurance coverage over time, an important consideration is the length of duration of spells of un-insurance and the capacity of this lack of coverage to lead to less efficient use of health care s

Methods: To improve the precision of survey estimates that characterize policy relevant population subgroups in a cost efficient manner, oversampling strategies are traditionally included as a core survey design component and implemented in the sample selection phase. This method of sampling involves the use of stratification to increase the precision of survey estimates for specific groups without incurring a major increase in overall sample size and survey costs. It also requires the availability of a sampling frame with the essential information to define the subgroups or strata and facilitate sample selection with differential rates. For population subgroups with characteristics that are static (gender/ race/ethnicity), the selection of an oversample that achieves targeted sample yields follows a conventional approach (Aday and Cornelius, 2006; Levy and Lemeshow, 1999; Kish, 1965). Alternatively, when the characteristic targeted for an oversample is subject to transitions over time (e.g. poverty status, employment, insurance coverage), the oversampling strategy is subject to much greater uncertainties in terms of achieving the desired sample size enhancements. When attempting to select an oversample of the uninsured, a substantial representation of the sample identified at the time of sample selection will have experienced a change in their coverage status at the time of interview. In essence, the effort is an attempt to hit a moving target. The greater the departure from a static characteristic, the more challenging the effort and the less certain the outcome in achieving sample size targets. Other obstacles that further limit the successful application of oversampling strategies relate to the level of availability of the key measures essential for the identification of the targeted population subgroup. Consequently, when attention is directed to an effort that attempts to increase the sample yield in a survey of individuals likely to be long term uninsured in the future, the operation is subject to both constraints at its inception: (1) the focus on a characteristic that is subject to change; and (2) a restricted set of available predictor measures available on a sampling frame.

Given the analytical and substantive importance of those individuals without health insurance coverage for extended periods of time, the development and specification of accurate models to predict the future likelihood of the occurrence of this event is highly desirable. At the outset, the specification of a clear definition of what constitutes the long term uninsured is critical. In this study, the ultimate objective was to

develop the best model to predict the set of adults under the age of 65 who are without any health insurance coverage for two consecutive calendar years. Given the low likelihood of long term uninsured children, the modeling effort was further restricted to adults between the ages of 18-64. In developing the prediction model, a core set of potential predispositional measures were identified that were applicable to health insurance take-up models and readily available from a screener interview. These included age, gender, race/ethnicity, health status, limitations in ability to work, marital status, education level, region, metropolitan statistical area (MSA) classification, presence of hospitalization, nativity in U.S., family size, poverty status and coverage status at time of screening. The measure of prior coverage distinguished whether the individual was covered at the time of the screening interview, and for the subset determined to be uninsured, the duration of time that had lapsed since the individual last had coverage (<6 months, 6 months- <1 year, 1 year- < 3 years, 3+ years).

Model Determination: In the final logistic regression model developed for predicting adults 18-64 likely to be continuously uninsured for two subsequent years, baseline health insurance status, race/ethnicity, marital status, education level, nativity, income and gender were determined to be significant predictors. The standard errors of all survey estimates and associated test statistics have been adjusted for the impact of clustering due to the complex multistage survey design and unequal weighting. When examining measures of model performance, the selected model exhibited the highest Pseudo R^2 (.228) and the lowest Akaike information criterion (AIC=4572.3). A receiver operating characteristic (ROC) analysis was also performed for each model, examining the area under the curve (AUC). The selected model also exhibited the highest AUC (.880).

Once the predictive model has been developed, additional analyses are necessary to identify the appropriate cut-off threshold in predicted probability for screening purposes to facilitate an oversample of this target population. To determine the operational cut-off point for each model, the predicted probabilities of being identified as continuously uninsured were determined for each sample individual based on the

underlying model specification. The predicted probability of being uninsured for two consecutive years in the future was derived from a transformation of an individual's predicted log odds from the respective prediction model. This set of predicted probabilities was then rank ordered by ascending values. The cut-off threshold to classify individuals as future long-term uninsured was determined to be 0.355, based on model performance with respect to predictive capacity, sensitivity, and specificity.

Evaluation Component: With specification completed, the models were evaluated in terms of accuracy in prediction. Examination of the performance of a model with the same data used to develop the best predictive model specification would risk contamination of the evaluation. Consequently, it was necessary to validate the models through application to an independent representative sample that characterizes the nation's health insurance coverage experience. This condition was satisfied through development of the prediction model using data from one specific MEPS longitudinal panel, and then applying the model to an independent MEPS longitudinal panel to assess model performance.

Model Performance: The evaluation of model performance examined predictive capacity, sensitivity and specificity, using the distinct predicted probability cutoff thresholds established with one MEPS longitudinal panel and applied to an independent sample. Using the alternative panel to assess performance, the final model correctly identified 54.9 % of those individuals in who were continuously uninsured. In addition, the model performed well with a specificity level of 94.1%. With respect to predictive capacity, the model correctly predicted 55.5 % of the target population (Table 1). The final criterion in model performance was directly focused on the expected sample necessary to support a 50 percent increase in sample yield, which would permit significant improvements in the precision of survey estimates which characterized the long term uninsured. Use of this metric facilitated an evaluation of the efficiency of a model based oversampling strategy to yield the targeted sample, standardizing the comparison in terms of sample size requirements under different model specifications. Using an assumption of a base sample requirement of 10,000 individuals aged 18-64 in a MEPS Panel responding for their entire two year period of eligibility in the survey, the required sample size necessary to

achieve a 50 percent sample size increase above the 1,173 expected long term uninsured survey participants was derived based on model-based oversampling with the selected model. This sample size specification calls for the inclusion of additional 587 individuals with the characteristic, resulting in overall target sample yield of 1,760 individuals who are long term uninsured in the survey. Results indicated that use of probabilistic models for oversampling purposes, to support a 50 percent increase in sample yield over a self-weighting design, permits the selection of the target sample of individuals who are continuously uninsured for 2 consecutive years in a cost-efficient manner. This methodology allows for an overall sample size specification for adults between the ages of 18-64 that is at least 25 percent lower than a design without access to the predictor variables from a screening interview such as the NHIS, or without application of oversampling techniques (Table 2).

Table 1: Model Performance, standardized sample of 10,000

Model 1: Logistic Model With All Significant Predictors

Actual: Insurance Status	Predicted as Long Term Uninsured	Not Predicted as Long Term Uninsured	Total
Long Term Uninsured	644 (54.9%: True +) (55.5%: correct prediction)	529 (45.1%:False -) (6.0%: incorrect prediction)	1,173 (11.7 %)
Some Coverage	517 (5.9 %: False +) (44.5% : incorrect prediction)	8,310 (94.1% : True -) (94.0%: correct prediction)	8,827 (88.3 %)
Total	1,161 (11.6%)	8,839 (88.4 %)	10,000 (100.0 %)

Table 2: Required sample size of adults 18-64 to yield sample of 1,760 individuals continuously without health insurance coverage over 2 years (50% increase).

	No Model-based oversample	Model based oversample: Model 1 –fully specified model	Model based oversample: Model 2-single baseline coverage measure	Model based oversample: Model 3-excludes baseline coverage measure

Expected sample yield of long term uninsured with baseline sample of 10,000 (no oversampling: equal probability sample design)	1,173	N.A.	N.A.	N.A.
Model prediction rate- % correct predictions	N.A.	55.5%	57.1%	38.8%
Ratio of predicted yield of long term uninsured to actual cases (inverse of prediction rate)	N.A.	1.80	1.75	2.58
Required sample augmentation for 587 additional cases (+50%)	5,000	1,058 (1.8 x 587)	1,028 (1.75 x 587)	1,512 (2.58 x 587)
Required overall sample size	15,000	11,058	11,028	11,512
95% Confidence Interval For sample size specification	14,880-15,120	10,969-11,147	10,939-11,117	11,420-11,604

Assumes base sample size of 10,000 individuals aged 18-64 in a MEPS Panel responding for their entire two year period of eligibility in the survey; equal probability sample design.

This examination of the performance of probabilistic models, to both identify and facilitate an oversample of the long term uninsured, revealed these model-based sampling methodologies to be effective statistical tools available for adoption in national health care surveys. This research effort sets the stage for additional modeling efforts to fine tune model parameters, examine their consistency for alternative time periods and to further improve upon predictive capacity. This study was limited by its dependence on the restricted set of predictor variables available from a screening interview. The inclusion of additional measures in a screening interview that improve predictive capacity (e.g. employment based offers of coverage, eligibility for public coverage) should also be considered. Research that informs the stability of results when focusing attention on alternative time intervals without coverage, such as a single year will also be beneficial. Future efforts will also be directed to the performance of the models for specific subgroups of the population classified by socio-demographic characteristics and income.

Furthermore, this framework also permits the development of models that predict the likelihood of maintaining continuous health insurance coverage for long time intervals into the future. The capacity to distinguish those individuals most likely to be continuously uninsured for long future time periods from those with much more limited spells could serve as valuable tool to assess future projections of coverage take-up.

5. Discussion

The federal statistical agencies responsible for the development and maintenance of core statistical and analytical health care related data systems that measure the "current state" of health and health care necessary to address high priority Presidential, Secretarial and departmental initiatives confront on-going challenges and trade-offs. Within HHS, the Department has established a Data Council to coordinate these health and human services data collection and analysis activities, to advance an integrated data collection strategy, to facilitate coordination of health data standards and privacy policy, and to address national health information infrastructure issues. The portfolio itself is quite dynamic and reflects considerable change to respond to Administration and Secretarial priorities. These changes often reflect enhanced capacity, improvements in data collection technologies and more focused and coordinated efforts that help ensure the availability of essential data to inform Secretarial priorities, and related departmental and program needs.

To complement assessments of the "current state" of health and health care, policymakers depend on model-based estimates of the "future state" under alternative demographic, economic and technological assumptions. These modeling efforts are major beneficiaries of the existing investments in health and health care data collection, and initiatives to ensure they yield efficient, well coordinated, integrated policy relevant datasets. However, they also place additional demands on data capacity, research, model development and statistical standards and rigor to better assess the impacts of revisions to existing health care policies.

In this paper, particular attention is given to the capacity of the Medical Expenditure Panel Survey to support such modeling efforts and ongoing strategies to advance their

utility for health policy formulation. To align these efforts with more conventional statistical analyses and provide metrics that serve to convey levels of uncertainty in the model outputs, a set of attributes that summarize the modeling process is advanced. In summary, this paper serves to emphasize the need for standards of data quality and statistical integrity in support of modeling and microsimulation efforts that are comparable to those developed for “current state” analyses. This is essential to ensure policymakers have a sound understanding of model assumptions, data limitations, and the level of uncertainty associated with these model-based estimates, prior to the implementation of a new initiative.

References

L. Aday and L. Cornelius. Designing and Conducting Health Surveys, Third Edition. San Francisco, CA: Jossey-Bass/Wiley, 2006.

Berk M, Monheit AC. The Concentration of Health Care Expenditures, Revisited. Health Affairs, 2001: 20:9-18.

Constance F. Citro and Eric A. Hanushek, Editors. Improving Information for Social Policy Decisions, The Uses of Microsimulation Modeling: Vol 1, Review and Recommendations. National Research Council, National Academy Press, Washington, D.C. 1991

S. B. Cohen. "The Utility of Probablistic Models to Identify Individuals with High Medical Expenditures". Journal of Economic and Social Measurement. 2005: Volume 30, No.2-3, 135-144.

S. B. Cohen. Design Strategies and Innovations in the Medical Expenditure Panel Survey. *Medical Care*. 2003; 4(7): 5-12.

S.B. Cohen. The Medical Expenditure Panel Survey: An Overview". *Effective Clinical Practice*. 2002; 6: 1-10.

S.B. Cohen,, T. Ezzati-Rice, and W. Yu. (2005) "Integrated Survey Designs: A Framework for Nonresponse Bias Reduction". Journal of Economic and Social Measurement 2005: Volume 30, No.2-3, 101-114.

S.B. Cohen, S. and W. Yu. "The Utility of Prediction Models to Oversample the Long Term Uninsured", Medical Care, 2008 forthcoming.

S. B. Cohen, S., T. Ezzati-Rice, and W. Yu. The Utility of Extended Longitudinal Profiles in Predicting Future Health Care Expenditures. *Medical Care.* 2006; 44 (5): (2006) 45-53.

S. B. Cohen, T. Ezzati-Rice, and W. Yu. The Utility of Probabilistic Models to Identify Individuals With High Medical Expenditures. *Journal of Economic and Social Measurement.* 2003; 30 (2-3): 135-144.

J. Hudson. Data from the 2002 NHEA-Aligned Medical Expenditure Panel Survey (MEPS): Interim Projected Medical Expenditures by Type of Service and Source of Payment for Age, Race and Sex through 2016. AHRQ Working Paper, June 2008.

Institute of Medicine. Health Insurance is a Family Matter. Washington, D.C.; The National Academies Press, 2002.

L. Kish. Survey Sampling. New York, NY: John Wiley and Sons, 1965

P. Levy and S. Lemeshow. Sampling of Populations: Methods and Applications, Third Edition. New York, NY: John Wiley and Sons, 1999.

Moeller JF, SB. Cohen, NA Mathiowetz, and L Wun. Regression-Based Sampling for Persons with High Health Expenditures: Evaluating Accuracy and Yield with the 1997 MEPS. Medical Care. 2003: 41(7):44-52.

Moeller, John F., S. B. Cohen, Ed Hock et al.."Projecting National Medical Expenditure Survey data: a framework for MEPS projections". Rockville (MD):Agency for Healthcare Research and Quality; 2002. MEPS Methodology Report No. 13.AHRQ Pub No. 02-0009.

Monheit, AC. Persistence in Health Expenditures in the Short Run: Prevalence and Consequences. Medical Care. 2003: 41(7):53-64.

T. M. Selden and J.S. Banthin. Health Care Expenditure Burdens among Elderly Adults: 1987 and 1996. Medical Care. 2003: 41(7): 13-22.

T. M .Selden and J. L. Hudson. Access to Care and Utilization among Children: Estimating the Effects of Public and Private Coverage. *Medical Care.* 2006; 44(5): 19-26.

A. Taylor, J. Cohen and S. Machlin. Being uninsured in 1996 compared to 1987: How has the experience of the uninsured changed over time? *Health Services Research*, 2001; 36(6-2): 16–31.

R. Weinick, S. Zuvekas and S. Drilea. Access to health care—sources and barriers, 1996. Rockville (MD): Agency for Health Care Policy and Research; 1997. *MEPS Research Findings No. 3*: AHCPR publication No. 98-000l. 1998.

[i] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey: Insurance/Employer Survey [AHRQ website]. Available at: http://www.meps.ahrq.gov/mepsweb/survey_comp/Insurance.jsp. Accessed July 10, 2008.

[ii] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey: Insurance/Employer Survey [AHRQ website].