Scenario Simulation Model: Data Sources and Database Construction

Supplement H to the Report:

Challenges and Alternatives for Employer Pay-or-Play Program Design: *An Implementation and Alternative Scenario Analysis of California's "Health Insurance Act of 2003" (SB 2)*

For the
California Health Care Foundation
and the
California Managed Risk Medical Insurance Board

Project Team Led by the
INSTITUTE FOR HEALTH POLICY SOLUTIONS

March 2005

Table of Contents

Acknowledgements

This supplement is part of a comprehensive report entitled *Challenges and Alternatives for Employer Pay-or-Play Program Design: An Implementation and Alternative Scenario Analysis of California's "Health Insurance Act of 2003" (SB 2).* The study that produced these reports was funded by a grant from the California HealthCare Foundation (CHCF), an independent philanthropy committed to improving California's health care delivery and financing systems.

Drs. Kanika Kapur and M. Susan Marquis of the RAND Corporation developed the database and estimation model, conducted the quantitative simulation analysis on the implications of SB 2 and alternative scenarios, and drafted this Supplement, as well as Supplement I describing the simulation process and Supplement A, setting out its results. The lead organizations for the overall study were the Institute for Health Policy Solutions and the RAND Corporation. Rick Curtis, of the Institute for Health Policy Solutions, served as the project director. Please see the main report for a detailed discussion of the study process.

The project team is grateful to Deborah Kelch, President of Kelch Associates and our liaison with the Foundation, who provided substantial assistance with the interview process and whose project management skills helped keep us on track. Last, but not least, Jill M. Yegian, Ph.D., and Marian R. Mulkey, M.P.H, M.P.P., of the California HealthCare Foundation provided analytic insights, helpful comments, and other support throughout the project. Without the assistance of the Foundation and its staff, this project would not have been possible.

Supplement H:

Scenario Simulation Model: Data Sources and Database Construction

Background

The California Health Insurance Act of 2003 (also known as SB 2) adopted a "pay-or-play" mandate aimed at reducing the size of the state's uninsured population. This law required employers over a certain size to either: (a) "pay" a fee to the state so that their workers and, for employers with 200 or more workers, dependents could be covered through a State Health Purchasing Program established under the Act, or (b) "play" by directly providing health coverage for specified workers and dependents. Although SB 2 was overturned by a narrow margin in a November 2004 referendum, the passage of legislation intended to expand employment-based coverage provides a unique opportunity to assess and to evaluate the implementation issues and challenges presented by a "pay-or-play" program.

Data sources

To characterize the employers and workers affected by SB 2 and to conduct our simulation, we require a database that includes comprehensive information on establishments and firms in California and their employees and dependents. Unfortunately, no such data exist. To address this data limitation, we created a database that describes employers and their employees in California in 2003 based on a number of sources including: data from the Employment Development Department (EDD) in California; the 2002 and 2003 surveys of private employers in California conducted by the Kaiser Family Foundation and the Health Research and Educational Trust (KFF/HRET); the 1997 Robert Wood Johnson Family Foundation (RWJF) Employer Health Insurance Survey; the 2001 panel of the Survey of Income and Program Participation (SIPP); and the 1997 Medical Expenditure Panel Survey.

Each of these data sources is necessary in the construction of our analytic employer-employee linked database. The EDD data provide current estimates of the number of firms and workers in California. However, these are aggregate data without any micro-level information on firm or workers characteristics. The KFF/HRET data and the RWJF data provide information on employers in California, including health plan offerings and worker composition. The KFF/HRET data have been recently collected, but they lack several crucial variables and need to be supplemented with RWJF data. The SIPP data provide comprehensive micro-level data on workers, but lack information on medical expenditures. The MEPS data are used to supplement the SIPP with medical expenditure data. We describe below how we used each of these sources in constructing the synthetic database of employers and employees in California and the non-working population.

Private employer database

We obtained special analyses from the EDD of the counts of private employers in California for eight employer size classes, with employer defined according to the SB 2 legislation to include all establishments in the state that are part of a single firm.[1] Separate counts were provided for multi-site businesses and single site businesses. EDD also provided us with total employment counts in each employer size class. We used the EDD data as benchmarks and developed new weights for the sample of employers included in the 1997 RWJF survey so that the weighted count of employers and employees in the adjusted RWJF sample matched the EDD benchmark totals. We then used these new weights to generate one record for each of the 1,013,506 private employers in California. We chose the 1997 RWJF sample to use as the base for generating employers because each record provides a count of the number of employees in the state as well as the number of employees in the country. It is this latter measure that is available from the population surveys, so we need this nationwide employment count to match workers and businesses.

We then linked information on characteristics of the business—including whether the business offered insurance, the share of employees enrolled in the group plans, its industry, whether it has union employees, and the age composition of the business-- by randomly selecting these characteristics from businesses sampled in the 2002 and 2003 KFF/HRET surveys of the same employer size and multi-site status as the business in our employer population. That is, we effectively reweight the KFF/HRET survey to match the EDD benchmarks. We used the pooled 2002 and 2003 surveys where possible. However, data on the age composition was only available from the 2002 survey and so we selected an age mix from the 2002 KFF/HRET sample controlling for business size, multi-site status, and industry.

We also selected characteristics of the plans offered from the 2002 KFF/HRET survey. We compared the 2002 and 2003 KFF/HRET surveys and found that the average value of most plan characteristics such as deductibles and copayments remained unchanged from 2002 to 2003. However, premiums had increased about 15 percent for all types of plans. We adjusted the 2002 premiums for this inflation. We chose to use the 2002 survey data for two reasons. First, the 2002 survey contains some information about the age of the workers that we used in matching workers to business; characteristics of plans offered and premiums are likely to differ in businesses that have different age profiles. Second, we impute actuarial values to the plans offered, and the 2002 KFF/HRET survey provided more measures of the plan characteristics that could be used for this purpose.

In order to obtain more detail about the distribution of characteristics of workers for our employer database, we again used the 1997 RWJF survey, which includes a more detailed age breakdown, as well as a distribution of workers by 4 wage categories[2], by gender, and by union affiliation. We controlled for size, multi-site status, whether offers insurance, whether any union

[1] The size classes provided by the EDD were 0-19, 20-49, 50-99, 100-199, 200-499, 500-999, 1000-4999, and 5000+.

[2] The categories in 2003 dollars were less than $8 per hour, $8-11.5 per hour, $11.5-17 per hour, and above $17 per hours. The 1997 survey dollars were inflated to 2003 dollars using the Consumer Price Index.

employees, and the share of young workers in matching these characteristics from the RWJF sample employers to employers in our constructed database.

The resultant database provides us with basic information about the business, some information about the composition of its workers on several socio-economic dimensions, and information about health insurance plans offered and worker enrollment decisions. Specifically we have information about the number of workers by age, gender, and whether in a union and by four age groups[3], four wage groups, and 5 health insurance enrollment decisions (in a non-offering business, offered but not enrolling, enrolling in single coverage, enrolling in two-party coverage, and enrolling in family coverage). These measures provide information about the number of workers in the business on each of several dimensions separately, but they don't provide information about the number of workers in each cell of the joint distribution of these characteristics. For instance, we know how many workers are in each of 4 wage groups and how many are in each of 4 age groups, but we do not know how many people are in each of the 16 groups defined by both wages and ages. We use the technique of Iterative Proportional Fitting to estimate the number of workers in each cell of the cross-classification of all of our dimensions (Causey, 2003). This technique uses information about the joint distribution of these characteristics in a population database and iteratively adjusts this distribution to match the marginal counts for the employer on each of the dimensions. The resultant solution gives us an estimate of the total number of workers in each of the cells described by the full cross-classification of the dimensions listed above. The joint distribution of these characteristics that we begin with is from the SIPP, which is the source of our population data and described in more detail later. We use the results from the iterative proportional fitting process to determine the mix of workers to select to populate each of our businesses.

Government employer database

The sources of data for government employers were EDD and the 1997 RWJF survey. The EDD provided information about the number of local government businesses by eight size groups and the number of workers in each class, and the total number of federal and state workers in California. The KFF/HRET data only sampled private employers and were not used here.

As we did in constructing the private employer database, we reweighted local government establishments included in the 1997 RWJF survey so that the number of businesses and the count of employees in each size group matched the EDD benchmark totals. We then created the 3,702 local government establishments. Each establishment has information about the distribution of characteristics of workers and information about health insurance plan offers from the 1997 RWJF establishment that represented it. We also use the technique of iterative proportional fitting described above to determine the mix of local government workers to assign to each local government establishment.

We created a single business entity for all federal workers in California and a single unit for state workers, because health insurance decisions are by and large made for the group as a whole. Information about the characteristics of federal and state workers and the health insurance offers were taken from the 1997 RWJF survey. We used EDD counts of federal and state workers as

[3] Under age 30, 30-40, 40-50 and older than 50.

the measure of the unit employment. Because we are representing federal and state governments as a single entities, all workers reporting that they are employed by the federal or state government in our population database are used to populate these two "employers", and we just adjust the weights in the population database to correspond to the EDD totals.

Actuarial value of plans offered

We represent the benefits of the plans that are offered by employers in our constructed database in terms of the actuarial value—the share of the total expenses incurred by the group that would be paid for by the plan. Insurance typically covers large medical bills more generously than small medical bills. Therefore, we have calculated the actuarial value of the plan benefits for workers in four groups classified according to spending levels based on the quartiles of the distribution of spending in the entire population, as well as an overall measure of the actuarial value. This latter measure is heavily weighted by coverage for large medical bills and does not represent what an average or median member of the group would expect to have covered.

We imputed the actuarial values to the plans offered by the private businesses in our database based on specific coverage provisions of the plans. The Actuarial Research Corporation previously constructed measures of the actuarial values of all plans that were offered by employers in the 1997 RWJF survey. This was done by simulating what each plan would pay for the spending reported by persons with group insurance in the 1987 Medical Expenditure Panel (adjusted to 1997 spending totals). Thus, the actuarial values of two plans differ only because of the benefit design and not because of characteristics of persons enrolled in the plan.

In order to account for changes in benefit design between 1997 and 2003, we estimate the relationship between actuarial values and benefit characteristics and then impute actuarial values based on current plan benefit designs. We fit regression models relating the actuarial values to specific features of the plans in the 1997 RWJF survey data, including the amount of the deductible, the amount of the copayment or coinsurance rate, whether coverage was provided for prescription drugs and mental health care, the copayment or coinsurance rate for these services, the out-of-network coinsurance rate for PPO and POS plans, and the maximum out-of-pocket expenditure for PPO and conventional plans.[4] We fit separate regressions for the four types of plan: HMOs, PPOs, POS, and conventional. Because actuarial values fall between 0 and 1, we fit a logistic model to the data to constrain the predictions to the appropriate range. This provided us with 20 regression models (the four quartiles of the actuarial values and the overall actuarial value for each of the four types of plan). We also have 4 sets of 5 empirical residual distributions—that is, the residuals from fitting the 5 equations for each of the 4 plan types. We then imputed actuarial values for each of the plans in the KFF/HRET survey by using the regression to predict the expected value of each of the 5 actuarial values associated with each plan and drawing randomly from the residual distribution to account for unobserved characteristics. We select jointly the 5 residuals to preserve the correlation in the distribution of errors across the actuarial values. That is, cases that have above average actuarial value for the lowest quartile of expenditures, given the observed characteristics of the plan, also have an above

[4] Dollar values in 1997 were inflated to current dollars using the medical services component of the Consumer Price Index. Since the actuarial value is a ratio of dollars, it is not adjusted.

average actuarial value for the other expenditure quartiles. The average correlation among errors for the HMO and conventional plans was .71 and it was .61 for the PPO and POS plans.

Population database

The primary data source for information about the population and its demographic, economic, and insurance characteristics is the 2001 SIPP panel. The SIPP provides information about work, insurance coverage, and access to group insurance for all family members—information that is necessary to accurately characterize who will be affected by SB 2. In addition, it provides demographic data and some information about health status. We selected all 8163 California respondents from the May 2002 cross-section of the panel. The May cross-section was chosen because it is the wave of data collection from the panel that includes questions about whether the employer offered insurance.

Respondents in the SIPP who are workers are used to populate each of the private and public employers with employees. We stratify private sector workers into one of 960 strata –defined by gender, whether union member, 4 wage groups, 4 age groups, 5 health insurance groups and 3 business size groups—fewer than 50 workers, 50-199 workers, and 200 or more workers. We randomly selected workers from these strata to populate each private employer, given the characteristics of the workers that we estimated as described earlier. Sampling of workers was done with replacement, so that a single worker may be assigned to more than one business, and also may be assigned more than once to a single business. In effect, this reweights the SIPP sample to accord with the characteristics of the worker population based on our employer database. Note that we assign people to businesses based on our best estimate of the mix of the workers in the businesses in 2003, so that changes in the composition of the workforce between the 2002 survey observations and the year we are trying to represent are accounted for.

We carried out a similar matching for local government workers, but used only two size classes (fewer than 50 and 50 or more). As noted earlier, we have only one employer representing federal employment and one employer representing state employment and reweight these observations in the SIPP to match the total employment counts provided by EDD. We compared the distribution of characteristics using the SIPP weights and in our reweighted database, and they were quite similar. That is, our development of the synthetic database did not substantially alter the estimates of the distribution of characteristics of workers.

Our employer measure of size is the number of workers employed in the state of California. The SIPP includes two size measures—the size of the establishment and the size of the firm nationwide. In order to obtain a measure of size of the employer that matches our concept and to obtain size classifications for the SIPP that match those we are using, we randomly selected a business in 1997 RWJF database (which includes measures of establishment size, statewide employment in the firm, and national employment) from the same nationwide firm and establishment size group, industry, and multi-site status as our SIPP worker and attributied the statewide employment size from the selected business to the worker.

Dependents are linked to workers, and so all dependents are reweighted according to the selection of the worker. We linked insured dependents to the worker that provides their group coverage. If the dependent does not have group coverage, we linked the dependent to the worker

in the largest business size group. In the case of ties, we linked the dependent with the male worker, since the majority of dual worker families provide coverage to dependents through the male's employer (Marquis and Kapur, 2004).

People in families that do not include any workers are retained in our data set with their SIPP sample weight.

For our simulation model, we require information on each family's marginal tax rates. We constructed tax-filing units using the SIPP data and information about the income ad exemptions for each unit. Marginal tax rates for each family were then estimated using the National Bureau of Economic Research TAXSIM model (Fennberg and Coutts 1993).

Expected health spending by the population

The SIPP collects information about the health status of respondents, but it does not include information about expected health spending. We use expected health spending by group members to characterize the risk of the group and will require expected spending to carry out our simulation analysis of the effects of SB 2 in forthcoming reports. So we have imputed health care spending to people in our SIPP sample based on health care spending reported by the sample in the 1997 MEPS. In an earlier study, ARC aged the health care spending reported by all privately insured respondents to the 1997 MEPS to the year 2002 using the CMS National Health Accounts as the benchmark. For each person in the MEPS, we have their spending, their age, and their self-reported health status. We stratify the MEPS sample into 30 groups –based on 6 age categories and the 5 levels of reported health status. We randomly select a level of health care spending for respondents in the SIPP sample from the appropriate stratum. This is a measure of actual spending in one period. In order to obtain a measure of expected spending for each person and the distribution of spending for each person, we repeat this process 5 times. Expected health care spending for the individual is then measured as the average over the 5 measures.

References:

Marquis, M.S. and K. Kapur, "Family Decision Making When Two Workers Are Offered Coverage", 2004, http://www.dol.gov/ebsa/.

Causey, B.D. "Dual System Estimation Based on Iterative Proportional Fitting", Census/SRD/RR-84/03, 2003, http://www.census.gov/srd/papers/pdf/rr84-03.pdf

Feenberg, D. and E. Coutts. 1993. "An Introduction to the TAXSIM Model." *Journal of Policy Analysis and Management* 12: 189-94.