ESTIMATION CONCERNS FOR FAMILY LEVEL ANALYSIS IN NATIONAL LONGITUDINAL HEALTH CARE SURVEYS,
Steven B.Cohen, Center for General Health Services Intramural Research, Agency for Health Care Policy and Research

Introduction

Designing a cross-sectional or one time survey contact to collect and analyze family level data is usually a straightforward process. Only those individuals and families in existence on the date of the interview or at a particular point in time are included in the study. Alternatively, national longitudinal surveys are affected by the changing structure of families over time. Throughout the reference period of the survey, families change their composition and new families are created or dissolved as a function of marriage, divorce, birth, death, separation, migration and institutionalization. Alternative definitions of family unit formation require rather complex weighting strategies to facilitate the derivation of national estimates of population parameters to characterize dynamic families over a specified time interval. In this paper, alternative estimation strategies for family level analysis of longitudinal data are examined, with particular applicability to data from the National Medical Expenditure Survey.

Data Collection Strategies for Family Level Characteristics in the National Medical Expenditure Survey

Traditionally, the primary analytical unit in most national longitudinal health care surveys is the person. As a consequence of the analytical requirements of these studies, which are directed to measuring and analyzing changes in the study population that occur over time, a longitudinal survey design is adopted to obtain data covering a fixed time interval. Longitudinal data collection strategies with multiple contacts (panel survey designs) are also adopted to reduce the impact of long recall periods on measurement error. Often, national longitudinal surveys are designed to satisfy both objectives. As an example, the National Medical Expenditure Survey (NMES), sponsored by the Agency for Health Care Policy and Research, was conducted to obtain national estimates of the health care utilization, expenditures, sources of payment, and health insurance coverage for the U.S. civilian non-institutionalized population for calendar year 1987. To reduce the deleterious impact of long recall periods on measurement error, data collection specifications required four separate interviews conducted with selected households at three to four month intervals over a fifteen month period. Furthermore, one of the primary analytical requirements of the survey was to measure changes in heath insurance coverage over the course of calendar year 1987.

The NMES data collection strategy for the household survey was motivated by the dual analytical goals of measuring health care utilization, expenditures, and health insurance coverage at both the individual and family levels for calendar year 1987 (Cohen, DiGaetano, and Waksberg, 1987). It should be noted that the reference to family level analyses in NMES will include consideration of families consisting of two or more individuals **and** consideration of single person households (families). Consequently, the concept of a family for estimation purposes in NMES overlaps with a household definition. To satisfy this goal, the sampling and data collection plan had to result in data for a probability sample of all persons who were civilian noninstitutionalized residents of the United States for all or part of 1987, and a probability sample of all families residing in the U.S. during all or part of 1987 that contained at least one civilian, noninstitutionalized person. To obtain accurate probability samples of these two groups required the development of inclusion rules to account for the many ways persons and families could enter and leave the target population during the reference period of the survey (Cox and Cohen, 1985).

The NMES probability sample of individuals was obtained in the following manner. First, a multi-stage national probability sample of dwelling units was selected (Cohen, DiGaetano and Waksberg, 1987). All civilian, noninstitutional residents of these sample dwelling units at the time of the first round of data collection were included in the NMES. In addition, persons entering the eligible population by birth or return from an institution, the military, or overseas, were included in the NMES when they joined a household containing one or more of the individuals initially sampled. These two groups of individuals, referred to as "key" individuals, constituted the core sample for all person level analysis in the NMES. Data for these sample persons were obtained for calendar year 1987 in four distinct interviews. Key sample persons who moved were interviewed at their new location whenever possible.

To facilitate data collection for families in NMES, the following family unit definition was operationalized. A family was defined to consist of individuals related by blood, marriage, adoption, or foster parent relationship. Unmarried college students between the ages of 17 to 22 who resided in a location distinct from their parents' residence were considered part of their parents' family. These individuals were only included in the survey when their parents' residence was sampled. A family that had a college student member living away from home would consist of two separate reporting units for purposes of interview administration. After data collection, the data obtained from the student reporting units would be linked to the parents reporting unit to facilitate the creation of family units. A reporting unit was defined to consist of individuals residing in the same dwelling unit who are related by blood, marriage, adoption, or foster parent relationship. Unrelated individuals were treated as distinct reporting units and defined as single person households.

The probability sample of families was obtained by including a family and all its members when it contained one or more "key" NMES sample participants. Over the course of the survey year, persons were added to the survey because of marriage to a "key" sample participant or as a consequence of joining the reporting unit of a relative who was a "key" sample participant. Except for those individuals joining "key" person's families in subsequent data collection rounds that were newborns, or individuals returning from the military, overseas residence, or institutions (who did not have an initial chance of

sample selection in NMES as a consequence of being out of scope for the survey during the first round of NMES interviews) the persons added to families after Round 1 were classified as "non-key". For nonkey family members, data were collected only for the time period in 1987 when they belonged to a family containing a key individual. The data collected for these nonkey persons are to be used only in constructing data aggregations needed for family level analyses (e.g., the annual expenditures incurred by the family for health care in 1987).

Alternative Strategies for Family Level Analysis in National Longitudinal Health Care Surveys

The two major complexities that affect the analysis of family level data in national longitudinal health care surveys are the changing structure of the family over time, and definitional problems in determining family continuation, dissolution, or formation. Throughout the reference period of a particular survey, families change their composition and new families are created by life events such as birth, death, marriage, divorce and separation. To cite a common event, suppose an initial family of a mother, father, and child at the start of the survey experiences a change due to the marriage and moving away of the child at a later point in time. The result of this family "split" yields a mother-father family and a child-spouse family. Should the mother-father family after the child's marriage be considered the same family as the initial larger unit, or be defined as a new family unit? A variation on the same theme is provided by Duncan and Hill (1985). Consider a household consisting of a husband, wife and three children at time 1. Between time 1 and 2, the couple separates, with the oldest and youngest children remaining in the original dwelling unit with their mother. Should the family consisting of the mother and two children remaining in the original dwelling unit be considered the same family as the larger unit, or defined as a new family? Explicit rules that govern family formation, dissolution, and continuity must be developed prior to the specification of an estimation strategy.

A broad spectrum of alternative strategies have been considered to facilitate family level analysis of data from national longitudinal health related surveys. These approaches range from a strictly dynamic treatment of families to those which totally ignore the changing composition of families over the course of the survey. Perhaps the most radical departure from a direct analysis of data at the family level is the strategy proposed by Duncan and Hill. Arguing that no single definition of a "longitudinal household" is appropriate for most analytic tasks, they contend that a superior alternative is the use of the individual as the unit of analysis, while attributing to each individual the characteristics of the household in which he or she lives. While such an approach has merit in determining the impact of changes in family composition at the person level with respect to an individual's annual health care utilization, expenditures, or health insurance coverage, its restriction to person level analysis is insufficient for surveys with an explicit requirement to yield family level estimates of health care parameters.
As noted, one of the explicit analytical objectives of the NMES was to derive family level estimates of health care parameters. More specifically, the following types of family level estimates of health care utilization, expenditure and insurance coverage measures will need to be derived from NMES data:
Mean Estimates
1. Mean number of ambulatory physician contacts per family;
2. Mean expenditures for ambulatory physician contacts per family.
Distributional Estimates
1. Percent of families with no ambulatory physician contacts in 1987.
2. Percent of families with out of pocket expenditures for ambulatory physician contacts above $2,000;
3. Percent of families with Medicaid coverage for at least one member at any time in 1987.
Since it is recognized that not all families will continue their existence throughout the entire calendar year, analysts will often attempt to control for family duration (e.g., derive separate estimates for stable families versus other families).
Whereas the family level estimates expressed in terms of population means are more amenable to estimation strategies that aggregate annual person level data to a single family, the need for distributional estimates imposes greater dependence on strategies that adjust for family formation and dissolution over the course of time. In the discussion that follows, the strengths and limitations of alternative family level estimation strategies with be examined in terms of satisfying the NMES requirements.

A. Cross-Sectional Treatment of Families
Perhaps the most straightforward of the remaining alternatives for family level analysis of longitudinal data is to consider a cross-sectional approach. Based on the representation of families for a specific time point during the survey reference period (often a year in duration), or for a given round of data collection, the entire longitudinal data profile of all associated individuals is attributed to these time-specific families. This treatment of attributing longitudinal characteristics to a family construct at a fixed point in time is comparable to the strategy adopted by the Census Bureau in the Current Population Survey (CPS), in determining annual household income. In this setting, the CPS combines the annual incomes from the preceding year for individuals who shared a household as of March in the current year, even when some household members were not present in the prior year, or when some household members for the preceding year have moved out by the March interview.

For longitudinal surveys comparable to the NMES, which collect health care data to derive annual estimates for relevant health care parameters, the time period covered by the initial interview or the final interview are the most viable choices for the time point or interval that defines families (DiGaetano and Brick, 1988). When the initial interview is selected as the time point, all the data for "key" members associated with the original family existing at this time point over the course of the year are aggregated to the family level to facilitate estimation. The original family serves as an "anchor" for all original "key" members in addition to newborns and other associated "key" persons

not eligible for sample selection at the time of the first interview. Much like the Current Population Survey treatment of individuals who do not reside in a selected household at the time of the interview, data for "non-key" persons that move into the set of original families over the course of the year are not included in the derivation of family level estimates . Since the "non-key" sample participants have already had a chance of selection for inclusion into the sample at the time of the initial interview, they are already represented in the original families by "key" survey participants.

The other alternative, that defines the set of families in a longitudinal survey at the time of the final interview, is subject to greater complexities with respect to estimation. Families that are defined at the final core interview consist of both "key" and "non-key" sample participants. These "non-key" participants have experienced multiple opportunities for inclusion in a longitudinal survey. A determination of their overall probability of selection requires additional information on their status at the time of the initial interview. Since data were collected for the "non-key" participants only for the period of time they were associated with "key" members of originally selected households, an explicit determination of their overall selection probability is problematic. A recommended approach for families consisting of both "key" and "non-key" sample participants at the final core interview is to determine the family status as a function of its reference person (the person who owns or rents the residence) (DiGaetano and Brick, 1988). When the reference person is a "key" sample participant, the family is to be included in the derivation of national estimates, with data aggregated from all of its members. Alternatively, families with a "non-key" reference person are to be excluded from all analyses. This strategy allows families that exist at the last round of data collection only one chance of selection in the survey. Adoption of a cross-sectional strategy for family level estimation and analysis has particular appeal in terms of its straightforward implementation and its suitability for deriving family estimates expressed in terms of population means. In addition, the consideration of a target population of families at a fixed point in time is consistent with the implicit assumption underlying standard post-stratification adjustments to sampling weights, both at the individual and the household level. Furthermore, the overall impact of linking sample persons to only one family and then aggregating annual person level attributes across these linked families imposes a consistency between national estimates of population totals for health care expenditure or utilization counts (the numerator component of the mean estimates) that are derived from person level or family level data. Using this approach, the mean estimates for families would converge with national utilization or expenditure estimates derived by taking the ratio of a person level estimate of the total health care expenditures or utilization counts for 1987, and a point in time estimate of the number of families in the U.S. for 1987. Development of family level weights under this model would be straightforward, with the family taking on the sampling weight of the "household/dwelling unit" or householder, which reflects its probability of selection into the sample.

Whereas the family level estimates expressed in terms of population means are more amenable to estimation strategies that aggregate annual person level data to a single family, the need for distributional estimates imposes greater dependence on strategies that adjust for family formation and dissolution over the course of time. More specifically, consider the following family level distributional estimate in NMES which attempts to measure the percent of families with annual out-of-pocket medical expenditures in excess of $2,000. By not adjusting for family formation and dissolution over the course of 1987, the cross-sectional approach treats all families existing at the time point that defines the cross-section as stable over the course of 1987.
Consequently, medical expenditures incurred by individuals while members of other families in 1987 will be attributed to the cross-sectional family. This approach will create an upward bias in the distributional estimate of the percent of families with annual expenditures above a specified threshold. Although it is possible to further distinguish cross-sectional families by those whose members are stable over the course of the survey reference period, in contrast with those whose members have experienced a change in family structure over time, this approach is not optimal for the derivation of distributional estimates.

B. Longitudinal Household Concepts

An alternative approach suggested for adoption in national longitudinal household surveys such as the Survey of Income and Program Participation (SIPP) conducted by the Bureau of the Census, is to restrict family level analyses to the subset that remain stable over the course of the survey reference period (Duncan and Hill, 1986; Citro et al., 1986). These families are often defined as "longitudinal households". When the reference period is limited to a year in duration, such as in NMES, the families that experience a change in composition may represent a small percentage of the total number of families in existence over the course of the year. When the time frame is lengthened, however, a substantial portion of the population will not be included in the derivation of family level estimates. For many policy relevant analyses, the households that experience a compositional change are often the most important to study.

To minimize the loss in representation of a family analysis strategy that is limited to longitudinal families, rules are established to maximize the number of originally sampled households that continue as "longitudinal" families throughout the course of the survey (Citro et al., 1986). One set of these rules are (Dicker and Casady, 1985):
1. two families will be linked in time if they have a common reference person and/or spouse of the reference person,
2. if a reference person/spouse split into two different households, the families with the most child months over time will be linked. This strategy allows the family with the most children to be defined as the longitudinal family,
3. if rule 2. results in a tie, the two families with the most family months will be linked. This rule attempts to define the longitudinal family by a comparison of household size, and
4. if the above rules cannot be implemented to make a determination of the families to be linked in time when they have a common reference person and/or spouse of

the reference person, a random linkage will be implemented.
Under this scenario, a family ends if it is reduced to a one person unit and begins when a one person unit becomes a two person or larger unit. As noted, these one person units would be included in family (household) level estimates required in NMES. Development of family level weights under this model would be straightforward, with the family taking on the sampling weight of the "household/dwelling unit" or householder, which reflects its probability of selection into the sample.

When the loss in representation is low, which is likely for household surveys with a reference period covering only a year, an estimation strategy for families that is restricted to the subset defined to be longitudinal is particularly attractive for the derivation of distributional estimates. Since all of the longitudinal families exist for a uniform period time, such as a year, annual distributional estimates derived by this approach can be viewed as standardized. While this standardization approach has appeal for the derivation of distributional estimates, the exclusion of a portion of the population that has experienced a fundamental change in household composition remains a serious concern for estimation. This exclusion may introduce significant bias in the resultant family level utilization and expenditure estimates, since these measures are often sensitive to changes in family composition.

C. Dynamic Families

A dynamic approach to family level analysis also requires an explicit set of rules for family continuity, dissolution, or formation. In the 1977 National Medical Care Expenditure Survey (Cox and Cohen, 1985) a family was defined to have changed composition when the household head or spouse departed. In this family unit framework, two new families were formed when the head or spouse moved out of an existing family unit and the original spawning family ceased to exist. Whenever there was a loss of the head or spouse due to death, institutionalization, or movement into the military, a new family was also formed and the original family ceased to exist. For changes in family composition concerning family members other than the household head or spouse, such as birth, death, movement out, or a member institutionalized or joining the military, the family was considered to be the same family, albeit with a different number of members. Obviously, alternative definitions of family formation, dissolution and continuity, which coincide with the longitudinal family definitions suggested for consideration for the Survey of Income and Program Participation (Citro et al., 1986), can be operationalized.

The advantage of this strategy is the inclusion of all families that are in existence over the timeframe for the survey in all analyses, and in the derivation of family level estimates. Consequently, the national estimates of family health care characteristics are representative of both stable families and families consisting of members that have experienced a change in family composition. An explicit estimation strategy for deriving family level estimates based on a dynamic model was developed for NMCES by Bentley and Folsom (1981). More specifically, the estimation scheme employed for dynamic families to derive annual mean estimates of health care utilization and expenditures is directed to the following population parameter:

$$\bar{Y} = \sum_{j=1}^{J} Y(j) / \sum_{j=1}^{J} E(j)$$

where Y(j) represents the health care utilization or expenditure experience for family j in 1987,

E(j) represents the fraction of days in 1987 that the family existed, and

J represents all families that ever existed in 1987.

The above specification insures convergence with utilization and expenditure totals for the nation, aggregated either across persons or families. Furthermore, the consideration of the factor, E(j), in the denominator of the population parameter, results in an annual average of the number of families existing over the course of the year, rather than a count of the number of families ever in existence over the course of the year. The denominator can also be interpreted as the number of family-years represented by all families that exist in 1987.
The estimator of this population parameter takes the same form, with the addition of sampling weights (W(j)), that have been adjusted for nonresponse and multiplicity:

$$\bar{Y} = \sum_{j} W(j)\, Y(j) / \sum_{j} W(j)\, E(j)$$

The implementation of this estimation strategy is dependent on a rather strict accounting of family transitions over time. Technically, the population of families existing on any day is potentially different from that existing on any other day due to the formation and dissolution of families over time. Person level data that cover the entire calendar year have to be appropriately allocated across time to the respective families the person is linked to over the course of the year. Furthermore, rather complex estimation strategies are necessary to consider, due to the need to appropriately reflect the multiple opportunities of selection experienced by the dynamic families over the course of the time period covered by the survey. In the NMES, where disproportionate sampling was employed to facilitate the oversampling of policy relevant groups such as blacks, Hispanics, the functionally impaired and the elderly, multiplicity adjustments to the sampling weights are less straight-forward.

The consideration of this estimation strategy is not without penalty. Greater data processing resources will need to be expended in order to allocate person level use and expenditure data across time to appropriate families, as contrasted with the cross-sectional and longitudinal family estimation strategies. As a consequence of missing dates associated with specific health care visits, additional bias will be introduced into the survey estimates. Furthermore, the multiplicity adjustment to the family level sampling weights will add greater variation to their distribution, which is often associated with a reduction in precision of survey estimates.

D. Recommended Approach

The analytical requirements for national longitudinal health care surveys comparable to the NMES necessitate the production of both mean **and** distributional estimates that characterize the health care experience of families.

Depending on the study timeframe, transitional families can represent a substantial portion of the target population. As a consequence, the dynamic family model is the recommended approach, which allows for the inclusion of both stable and transitional families in the derivation of national family level health care estimates. Given the complexities associated with the implementation of this strategy, and the need to provide timely national health care estimates at the family level, an incremental approach to family level estimation is proposed. More specifically, the following sequence defines the NMES family level estimation tasks associated with an incremental approach:

Phase One The cross-sectional strategy for family level estimation and analysis is initially adopted for estimates expressed in terms of population means. Family level weights are derived based on the sampling weight of the "household/dwelling unit" or reference person, and poststratified to national population estimates of households obtained from the Current Population Survey (Bureau of the Census) for the time point that defines the cross-section. No distributional estimates are produced at this stage.

Phase Two An explicit set of rules for family continuity, dissolution, and formation are developed to support NMES family level analyses. Based on these rules, families are defined to be longitudinal or transitional over the course of 1987. As noted, families that consist of both "key" and "non-key" sample participants introduce additional complexities with respect to estimation. For families consisting of both "key" and "non-key" sample participants at any time during 1987, family eligibility for estimation is determined as a function of its reference person (the person who owns or rents the residence). When the reference person is a "key" sample participant, the family is to be included in the derivation of national estimates, with data aggregated from all of its members. Alternatively, families with a "non-key" reference person are to be excluded from all analyses. This modified estimation strategy for the dynamic family model allows families that exist during the year only one chance of selection in the survey and obviates the need for a multiplicity adjustment. Preliminary distributional family level estimates are derived only for families defined to be longitudinal, using family level weights based on the sampling weight of the "household/dwelling unit" or reference person.

Phase Three

Person level data that cover the entire calendar year are allocated across time to the respective families the person is linked to over the course of the year. Family level weights based on the sampling weight of the "household/dwelling unit" or reference person are adjusted for the fraction of days in 1987 that the family was in existence. Using the dynamic family model, both longitudinal and transitional families are included in mean **and** distributional estimates that characterize the national health care experience of families for calendar year 1987.

Summary

In this paper, the estimation concerns associated with family level analysis in national longitudinal health care surveys are examined. The advantages and limitations of alternative analytical strategies for family level analysis are discussed, with specific references to examples from the 1977 National Medical Care Expenditure Survey, the 1987 National Medical Expenditure Survey, and the Survey of Income and Program Participation. Particular attention has been given to three alternative family construct specifications: cross-sectional families, longitudinal families and dynamic families.

Adoption of a cross-sectional strategy for family level estimation and analysis has particular appeal in terms of its straightforward implementation and its suitability for deriving family estimates expressed in terms of population means. However, the need for distributional estimates imposes greater dependence on alternative strategies that adjust for family formation and dissolution over the course of time. When the loss in representation is low, an estimation strategy for families that is restricted to the subset defined to be longitudinal is often implemented for the derivation of distributional estimates. While this standardization approach has appeal for the derivation of distributional estimates, the exclusion of a portion of the population that has experienced a fundamental change in household composition may introduce significant bias in the resultant family level utilization and expenditure estimates. Consequently, when analytical requirements necessitate the production of both mean **and** distributional estimates that characterize the health care experience of families, a dynamic family model is the recommended approach.

References

Bentley, B.S. and Folsom, R.E. (1981). Family Unit Analysis Weighting Methodology for the National Medical Care Expenditure Survey. RTI Final Report No. 1320-11F, Contract No. HRA 230-76-0268, Available from the Agency for Health Care Policy and Research.

Citro, C. F. Hernandez, D. J. and Moorman, J. E. (1986). Longitudinal Household Concepts in SIPP. Proceedings of the American Statistical Association, Survey Research Methods Section.

Cohen, S.B., DiGaetano, R. and Waksberg, J. (1987). Sample Design of the Household Component of the National Medical Expenditure Survey. Proceedings of the American Statistical Association, Survey Research Methods Section.

Cohen, S. B. (1982). Family Level Analysis in the National Medical Care Expenditure Survey. Proceedings of the American Statistical Association, Survey Research Methods Section, 561-566.

Cox, B. G. and Cohen, S. B. (1985). Methodological Issues for Health Care Surveys. Marcel Dekker, Inc. New York, New York.

Dicker, M. and Casady, R. J.(1985). An Introductory Discussion of Issues in the Methodology of Longitudinal Family Surveys. National Center for Heath Statistics Working Paper.

DiGaetano, R. and Brick, M. (1988). Considering Sample Weights for Families and Unrelated Individuals. NMES Memorandum 2-01700, Westat, Inc., Rockville, MD.

Duncan, G. J. and Hill, M. S. (1985). Conceptions of Longitudinal Households, Fertile or Futile? Journal of Economic and Social Measurement, 13: 361-375.

McMillen, D. B. and Herriot, R. A. (1984). Towards a Longitudinal Definition of Households. Survey of Income and Program Participation Working Paper Series No. 8402. Bureau of the Census.

Presented at the Annual Meetings of the American Statistical Association, Anaheim, California, August, 1990.
The views in this paper are those of the author and no official endorsement by the Agency for Health Care Policy and Research or the Department of Health and Human Services is intended or should be inferred.