Appendix A
Data Construction

The data for this dissertation are drawn from the Care after the Onset of Serious Illness (COSI) data. This appendix describes in some detail the construction of COSI and of the subset analyzed for this dissertation. COSI seeks to exploit the potential for multiple levels of analysis made possible by the existence of many different electronic data sets. These data include Medicare claims regarding patients' inpatient and outpatient health care use, and regarding medical providers, as well as other data sources such as the Census, America Hospital Association data, and the like. This appendix also serves as a record of the rationale, both empiric and theoretic, for a number of the design decisions that must be made in the development of such a dataset.

The Appendix is structured as follows. First we review the general structure of COSI. I define the precise criteria by which cases of interest were defined. Next I define the ways in which spouses are identified. Given this information, it is then possible to precisely define the cohort of interest to this dissertation. Thereafter I define a number of control variables that were used. After providing some information on the initial hospitalizations, I conclude with a discussion of the linkages to external data sources that were used. The general data structure of COSI is schematized in Figure A.1.

A.1 Empanelment

A.1.1 General Data Source

The core data used to develop COSI are the 1993 inpatient hospitalization records from the Health Care Financing Administration's Medicare program. These records, embodied in the "MedPAR" file, represent a complete enumeration of the final adjudication of all claims for hospitalizations filed by all Medicare beneficiaries for any hospitalizations or parts thereof occurring at any time during 1993. Medicare claims and enrollment data capture 96% of the American 65+ population. (Hatten 1980)

COSI was constructed to have two components. These include a series of disease-specific cohorts, which contain all individuals who met empanelment criteria for a given disease in 1993. The second consists of a unified cohort of individuals who met any empanelment criteria in 1993; naturally, in a large cohort, there were some individuals who independently qualified for two or more disease-specific cohorts. The COSI unified cohort is an individual-level cohort; individuals who qualified for more than one disease-specific cohort entered the unified cohort only once, as described below. My dissertation examined a subset of the unified cohort; the sub-cohorts are described as needed in this paper for clarity.

A.1.2 Selection of Initial Conditions and their Operationalization

The COSI project focuses on the longitudinal health care course of patients who have the new onset of a serious disease in 1993. The year was chosen arbitrarily, far enough back that substantial follow-up would be available, sufficiently recent so as to take advantage of Medicare's significant improvements in the quality of its electronic records at the end of the 1980's. Our objective was to construct an inception cohort of patients newly diagnosed with one of several serious illnesses, based on exploitation of hospital records.

Diseases chosen were chosen so as to meet the following criteria:

1. SEVERITY: the disease had a reasonably high probability of substantial mortality or morbidity
2. ACUTENESS OF ONSET: the natural history of the disease is marked by a point of threshold increase in its manifestations
3. LIKELIHOOD OF HOSPITALIZATION: the threshold increase in the manifestation of the disease is *very* likely to result in an acute hospitalization regardless of other characteristics of the patient
4. RELIABILITY OF DETECTION: the disease needed to be detectable in the claims with both high sensitivity and specificity
5. EPIDEMIOLOGIC SIGNIFICANCE: the disease must account for a reasonable burden of disease in the population
6. THEORETICAL INTEREST: the diseases should be sufficiently different in their natural histories so as allow generalizations in other domains.

The following diagnoses were assembled:

- lung cancer
- colorectal cancer
- leukemia
- lymphoma
- pancreatic cancer
- urinary cancer
- liver/biliary cancer
- CNS cancer
- head/neck cancer
- stroke
- congestive heart failure
- acute myocardial infarction
- hip fracture

The following were considered but ruled out:

- breast cancer
- prostate cancer
- chronic obstructive pulmonary disease
- dementia
- trauma

To define such cases, we relied on ICD-9-CM codes in the hospital claims; this required selecting ICD-9-CM codes that represent these diseases. In many cases, multiple definitions were identified in the published literature. We then chose the definitions with the best published empirical performance, when available. In the absence of comparative performance data, we chose definitions that most coincided with the best accepted research in the subfield, for example, the SEER definition for cancers, or the Cooperative Cardiac Project definition for acute myocardial infarction. Table A.1 provides the actual definitions, as well as references and comments.

After defining the ICD-9 codes to be used to identify a diagnosis, the appropriate exclusions needed to be implemented in order to capture *incident* cases. Prior detailed empirical work provided guidance here. Research examining the Medicare/SEER linked data demonstrated that for lung, colon, and esophageal cancers, three years of lookback in the Medicare claims was adequate to eliminate prevalent cases. (McBean, Warren and Babish 1994) That is, if an individual had not been *hospitalized* in the prior three years before the putative index hospitalization for onset of his/her serious disease, it was very likely that they had *never* previously been hospitalized for the disease. As such, hospitalizations for a disease with no similar hospitalizations in the past 3 years served as our operational definition of an incident index hospitalization. This makes conceptual sense as well — it is extraordinarily unlikely for an individual to have multiple independent primary tumors in the same organ within three years. That same research group, however, demonstrated that even

6 years of lookback were unlikely to clear all prevalent cases of breast and prostate cancer. Since onset of these disease could not then be identified, such conditions were removed from consideration from our cohort.

Hospital claims records can have up to 11 diagnoses, one "primary position" for the disease most responsible for the hospitalization, and 10 in "secondary positions" for diseases which contributed to the stay. Once the look-back had been defined, it was necessary to decide whether (1) to require that onset of a disease be defined as only those hospitalizations for which the disease of interest was noted to be the primary cause of hospitalization, or (2) to also accept as index hospitalizations those hospitalizations where the disease was noted to be contributory to the patient's hospitalization. As Table A.2 demonstrates, these differences could lead to substantial differences in the apparent incidence of the disease. Here, we relied on three sources: precedents, alternative epidemiological data, and clinical experience. Table A.3 demonstrates the comparison between our final definitions and existing epidemiologic data based on sources other than the Medicare claims, where such data exist.

For cancer patients, we accepted as the index hospitalization any hospitalization during 1993 which indicated a cancer diagnosis as defined in Table A.1 in any position, as long as the patient had never previously had a hospitalization where this cancer diagnosis had been noted. Naturally, the exclusion criteria were disease-specific; a diagnosis of acute myocardial infarction in 1992 did not prevent a patient newly hospitalized in 1993 with a lung cancer diagnosis from entering our

cohort. A similar line of reasoning went into the choice of methods for congestive heart failure.

Unlike most cancers or CHF, it is quite possible for a patient to have more than one incident stroke, myocardial infarction, or hip fracture. Therefore, the use of a look-back to exclude prevalent cases is less satisfying—it may lead to an inappropriately *healthy* selection bias by excluding those with multiple cardiac or intracranial events or fractures. However, for stroke and MI, our clinical experience and the past practice of other researchers both agreed that for an individual having an incident event, that event will be their primary diagnosis for that index hospitalization. This choice is reinforced by the distinction in the ICD-9-CM, exploited by the Cooperative Cardiovascular Project, between initial visits for a an acute MI and follow-up care visits. (Krumholz et al. 1998) In the case of hip fractures, it was *not* necessary to restrict the definition to only those cases where the diagnosis was in the primary position.

In the sole case of congestive heart failure, because of the commonness of the disease, we took a simple random 1 in 3 sample of all detected incident CHF patients to enter into our study cohort.

Cohort construction to this point had allowed individuals having index hospitalizations in more than one disease within 1993 that met our enrollment criteria to be enrolled multiple times. This was done to allow complete enumerations during disease-specific analyses to be maintained. During full-cohort analyses, patients were entered into COSI under whatever the temporally first diagnosis was. That is, if a

patient had hospitalization in 2/93 that met criteria for an index hospitalization for MI, and then had a second hospitalization in 5/93 that met criteria for pancreatic cancer, that individual would appear in both the MI and pancreatic cancer sub-groups, and would appear once in the unified COSI analytic cohort under an MI diagnosis. This occurred in 7.9% of unique individuals; see Table A.4. In cases where the diagnoses for a single claim qualified a patient simultaneously for more than one diagnosis, the patient was empanelled separately into each of the disease-specific cohorts. They were empanelled into the combined cohort under the cancer diagnosis preferentially if there was a cancer and a non-cancer that both qualified simultaneously. 54,037 (4.2%) patients qualified for 2 cohorts on the same day; 495 (0.04%) qualified for 3 on the same day, and 2 patients qualified for 4 cohorts on the same day.

A.1.3 Other Empanelment Issues: Exclusion of Other Diagnoses

As mentioned above, breast cancer, prostate cancer, chronic obstructive pulmonary disease, dementia, and still other diseases and conditions were considered as other COSI-empanelling diagnoses. Breast and prostate cancer were excluded based on the SEER/Medicare evidence that an accurate index hospitalization could not be identified in the sense that (1) not all patients with the disease are hospitalized for this condition, and (2) when they are, the hospitalization is not usually very near the diagnosis. (McBean, Warren and Babish 1994) Clinical experience suggested that the natural histories of C.O.P.D. and dementia were stories of gradual worsening with

occasional hospitalizations only in some situations; moreover, claims data are not sensitive for these conditions.

A.1.4 Other Empanelment Issues: Exclusions Based Only on Inpatient Acute Care Hospital Claims

While the bulk of MedPAR claims are for acute hospitalizations (11,307,844 of 12,709,289 or 96%), a number of other types of "hospitals" are included in the claims. These include primarily skilled nursing facilities, but also some so-called "long-term acute care facilities." We decided to use only claims from acute hospitalization for empanelment index hospitalizations. This allowed greater consistency with previous work, and accorded with our mental model of what an index hospitalization "should" be. To maintain consistency, a patient with a previous hospitalization at a non-acute-care hospital for a COSI-defining diagnosis could still be included in the cohort; that is, *only acute care hospitalizations were used to exclude prevalent cases*. Very few individuals would have been excluded from the cohort had we relaxed this restriction.

Similarly, claims for outpatient care (either Part A or Part B) were *not* used to detect the onset of disease, nor were they used to exclude patients as having prevalent rather than incident disease. This was based on our judgement. The natural history of many of these diseases – particularly the cancers – includes a premorbid period where disease burden is accumulating – usually with few or no symptoms. The initial manifestation of the disease is usually associated with a threshold change in the

intensity of symptomotology, often requiring an index hospitalization for diagnosis and treatment. As such, it seems to us that an individual with disease noted exclusively in the outpatient claims is likely to represent a different sort of disease than someone who has needed to be hospitalized – a lower level of severity, not yet "severe illness". This was, it must be emphasized, first and foremost a "judgement call" about our preferences as regards the scope of illness of interest to us in this project. Also, given how much less was known about the detection of disease in the outpatient claims, and what exactly that signifies, we did not feel that the methodological tools were yet available to explore this lower severity disease. (This continues to be an active line of research for a number of scholars. See, for example, (Cooper et al. 1999; Du et al. 2000; Klabunde et al. 2000; McClellan, Roghmann and Schilling 1998; Warren et al. 1999))

A.1.5 Other Empanelment Issues: Definition of a Single Hospitalization

Hospitals bill HCFA quarterly for the care of their patients. As such, a patient whose stay spans two (or more) billing cycles may have two (or more) separate claims filed for the same hospitalization. As others have done, (Palmer et al. n.d. (?1994?)) we identified any patient with multiple bills on which the discharge date of one was the same as or immediately preceded the admission data of the second. If these claims were filed from the same hospital, we declared them a billing artifact that represented in truth a single hospitalization for the patient.

A.1.6 Other Empanelment Issues: Minimal Data Completeness Restrictions

In order to be empanelled into our cohort we required the following minimal data integrity checks from the claims: (1) valid birth date, in order to impose the age restrictions; (2) a valid admission date; (3) some valid ICD-9-CM codes that met our enrollment criteria. The presence of other data errors on a claim (e.g. some invalid ICD-9-CM codes in other diagnostic fields) did *not* exclude a claim from empanelment. Therefore, there were some remaining data errors in the claims which required exclusion of claims because of incoherent dates (e.g death dates reported later than the day on which we took delivery of final mortality follow-up or before the admission date), missing sex, or race. Final cohort size with adequate minimum data was 1,231,894 unique individuals; this represents 99.19% of the 1,241,935 unique individuals initially screened for possible inclusion in cohort (*i.e.*, 0.8% were excluded due to miscellaneous data impurities).

A.1.7 Other Empanelment Issues: Geographic Restrictions

At this point we have not imposed any geographic restrictions. Thus the cohort includes individuals in Puerto Rico, Guam, and the U.S. Virgin Islands, and other miscellaneous territories. As supplementary material such as Census data is not in general available for these areas outside the 50 US states and the District of Columbia, we excluded such cases from many analyses. 1,221,153 probands lived within the 50 United States and the District of Columbia.

A.2 Finding Spouses

A.2.1 Overview

We have previously published a method to detect the marital status of many Medicare beneficiaries based on information latent in their claims. This method allows us to uniquely link individuals to their spouse. The details of this method have been described (Iwashyna et al. 2001; Iwashyna et al. 1998) and discussed elsewhere. (Iwashyna et al. 2000; Kestenbaum 2000) Briefly, it is has long been known that some married and widowed individuals file Medicare claims under a Health Insurance Claim number (HIC) that consists of their spouse's Social Security Number and a code indicating that the filing individual is a "dependent beneficiary" rather than a primary beneficiary. Moreover, individuals can change the HICs they use, particularly when their spouse dies. This has necessitated the use of "cross-reference files" when linking an individual across multiple years or types of HCFA data. What we noted previously was that while this causes hassles for constructing individual-level longitudinal data sets, (Parente et al. 1995) it also permits the development – using only information present in the claims – of longitudinal couple-level data sets.

In order to apply these methods, we could not restrict ourselves exclusively to the MedPAR files and other utilization-based claims. After all, a spouse might exist but not use any health care during our year of interest. However, HCFA also maintains an enrollment database. The 1993 Denominator file contains basic

identifying information on the entire Medicare population during 1993 – that is, it contains information on all individuals who were enrolled in Medicare at any point in 1993, regardless of whether or not they actually filed a claim. The enrolled population has been previously shown to closely approximate the population of all Americans age 65 and above. (Hatten 1980; Kestenbaum 1992)

We received from HCFA the 1993 Denominator file of 38,212,735 records, with mortality follow-up *for the entire Medicare population* through July 6, 1999, and a cross-reference file as of January 6, 1999. These mortality and cross-reference files were the most recent available at the time of this particular data request which is part of an ongoing research effort. We used this data to develop a list of all detectable husband-wife pairs as of 1993, where both were enrolled in Medicare at some point during that year. After doing so, we would "look-up" the spouse of our COSI-cohort members, matching them if possible. This allowed us to determine who was married (so far as we could detect) on Jan.1, 1993. We then took the unmatched population, and probed the Medicare data to see if they ever had had a spouse, which would allow us to not merely designate them a widow or widower, but to know for precisely how long they had been in such a status. The methodological importance of such precise information on the date of widowhood has been demonstrated by others. (Korenman, Goldman and Fu 1997)

A.2.2 Finding Spouses in the Denominator File: Direct Matching

The 1993 Denominator file contains 38,209,888 unique individuals, 32,180,588 of whom were at least age 65 as of January 1, 1993. Altogether, there were a total of 36,915,227 Health Insurance Claim numbers (HICs) in the cross-reference file used by these 32 million individuals. Of those 65+ in the file, 10,110,008 have died by our follow-up on July 6, 1999. (All further information is confined exclusively to the 65 and above population unless explicitly stated.)

A total of 5,188,168 individuals (2,594,084 couples) could be linked directly in the 1993 Denominator file using direct linkage of primary and secondary recipients, as described in Method 1 in our 1998 *Demography* article. As was known, a single primary beneficiary may have more than one dependent beneficiary; this occurred 36,464 times. In these tables, only the most recent marriage is counted. Using the cross-reference file, we were able to link an additional 3,438,274 individuals (1,719,137 couples) who were both alive as of January 1, 1993; this corresponds to the "Method 2" of our original manuscript.

However, for the purposes of this dissertation, these "Method 2" couples could not be analyzed. I used only "Method 1" couples – those couples who could be prospectively identified as married without reference to the cross-reference file. The reason is this. A major way in which a couples is detected for Method 2 is that the higher earning husband dies first. The wife can then be linked to the husband as a result of the change in benefits that occurs. However, this constitutes selection on our variable of interest. That is, there is no comparison population of interest – even the

population of wives whose husbands are still alive are wives whose husbands are going to die within a few years. Thus the hazard ratios are of unclear significance for the broader question of interest – the relationship between the death of one spouse and the mortality of the other. We therefore excluded all Method 2 couples from our analyses for this dissertation.

A.2.3 Finding Predeceased Spouses from our Cohort: Hypothetical HIC Generation

At this point in the construction of COSI, we had identified about half of the men and one-fifth of the women in our cohort as married. However, the remainder of the cohort was mix of married, single, widowed, and divorced individuals. Among the women, in particular, we expected there were large numbers of widows who were receiving benefits based on their husbands' income (with the husband having died before 1/1/93). In order to determine this, in February, 2000, we sent a file to HCFA which contained a set of hypothetical HICs. That is, we took all the HICs ever used by any cohort member not matched to a spouse alive in 1993 by the process using the Denominator file. We then changed the BICs to their reciprocal -- where we had a BIC indicating a dependent spouse or widow, we created a HIC indicating a primary beneficiary, and vice versa. After excluding duplicates and already-detected individuals, we sent this list of 4,963,942 “hypothetical” HICs to HCFA and asked them to identify any who had ever been alive, and their dates of death.

This process allowed us to identify 271,105 widows and 54,700 widowers. Further, because some spouses became eligible for Medicare in the interval between the end of 1993 and our sending of this data, we detected 31,425 new married couples among our cohort. Thus, the final distribution of our cohort by marital status as of Jan. 1, 1993, is shown in Table A.5. Because the dates of the transitions from married to widowed are know precisely, these dates are recorded in the data set. Therefore, for cohort members with multiple index diagnoses, we can determine their precise marital status on the day of admission *for the disease under consideration* at that time in disease-specific analyses. Moreover, we can explore in detail the time course of effects of marital status transitions. These widowed by definition survived the higher hazard of death likely associated with bereavement – they had to survive until bereavement. As such, this selects for "healthy" survivors, and cross-sectional comparisons would likely *under*estimate the true difference between the married and the unmarried in a prospective cohort. We examine this population of all widows, without regard to time since bereavement, only in the first part of Chapter 2 and for Chapter 4, on the impact of marital status on health care utilization.

We cannot, using these methods, detect *new* marriages among the elderly – a very rare phenomenon. (In 1990, among the previously widowed, the rate was 1.7 marriages per 1,000 for elderly women, and 14.0 per thousand for elderly men; thus, it is unlikely that more than a few percent of our widowed sample may have remarried. (Clarke 1995)) Likewise, we cannot detect cohabiting couples, a similarly

very rare phenomenon at this point in time in this population. (Bumpass and Sweet 1989; Chevan 1996)

A.2.4 Dealing with Divorcees

These data may be contaminated by couples who are divorced. Some of the members of these former couples may qualify for dependent spousal benefits, although the restrictions are quite strenuous. While fewer than 5.7% of the elderly are divorced, (U.S. Bureau of the Census 1996) some of these former couples may contribute to the overestimation of our detection efficacy. The precise details of how one qualifies for divorcee benefits are arcane, and we were not confident of our ability to exclude all divorcees. Instead, I required that married couples have the same mailing address ZIP code at the time of the proband's admission, which occurs in 87.24% of detectably married ("Method 1" or "Method 2") cases. The married but not coresiding were excluded from all analyses.

A.3 Variable Definitions

A.3.1 Defining the Death Date

Death dates were obtained from the highly accurate Vital Status file of the Health Care Administration as of July 6,1999. This file is updated regularly from the Social Security Administration. This file has been shown to be highly accurate, although there are known defects in the detection of death of certain very old widows.

(Kestenbaum 1992) Table A.6 shows the rates of death for all cohort members by the end of 1997, after at least 3 years of follow-up.

A.3.2 Expanding the Racial and Ethnic Coding Based on Name Algorithms for Hispanicity and Asian Origin

Medicare data have certain well-known limitations with respect to their racial classification system, and the race codes provided in the claims can only be reliably used for white/non-white comparisons. (Arday et al. 2000; Lauderdale and Goldberg 1996) However, our data included the beneficiary names. As such, we were able to apply well-validated algorithms for identifying Hispanic and Asian-American ethnicities, substantially improving the adequacy of the racial/ethnic classification system we can use here. (Lauderdale and Kestenbaum 2000; Word and Perkins Jr. 1996) As such, "white" and "black" as used in this manuscript refer to non-Hispanic white and non-Hispanic black; the shorter words are used for expositional convenience. 28,719 probands had their race codes reassigned by the algorithms used. As expected given the geographical racial distribution, (Sandefur et al. 2001) 85.5% of these reclassified Hispanic and Asian-Americans lived in the states of Arizona, California, Florida, Hawaii, Illinois, Massachusetts, New Jersey, New Mexico, New York and Texas.

A.3.3 Developing Comorbidity Measures

In order to make valid mortality comparisons between groups, differences in health at baseline must be taken into account. One fruitful way to operationalize "health" for such purposes is the notion of comorbidity burden. A comorbidity is a chronic disease of substantial mortality, morbidity, or management burden. The number of such comorbid conditions a patient has are often aggregated into comorbidity index to provide a simple scalar measure.

Among the most popular comorbidity indices in claims data research are those based on the work of Mary Charlson and her collaborators,(Charlson et al. 1987) particularly as implemented in the ICD-9-CM codes for computerized use.(D'Hoore, Sicotte and Tilquin 1993; Deyo, Cherkin and Ciol 1992; Romano, Roos and Jollis 1993a) While several alternative risk adjustment approaches have also been published,(Brailer et al. 1996; DesHarnais et al. 1990; Elixhauser et al. 1998; Fowles et al. 1996; Iezzoni 1997; Iezzoni et al. 1994; Kuykendall et al. 1995; Schwartz et al. 1996; Starfield et al. 1991; Weiner et al. 1991) the Charlson method is extremely popular and has been used extensively.(Christakis and Escarce 1996; D'Hoore, Sicotte and Tilquin 1993; D'Hoore, Bouckaert and Tilquin 1996; Iwashyna et al. 1998; Roos et al. 1989) Direct comparisons between these alternative scales are relatively rare, and the choice of the Charlson index is somewhat arbitrary.(Ghali et al. 1996; Hughes et al. 1996; Romano, Roos and Jollis 1993b; Roos, Sharp and Cohen 1991) On the whole, these indices have been developed for the prediction of *mortality* following *hospitalization*, a situation quite similar to the uses to which we will put

them. We have previously shown that statistically and empirically significant improvements in the prediction of mortality were obtained by incorporating alternative sources of data —particularly two years of inpatient lookback combined with one year of outpatient and auxiliary claims lookback — but only if indices derived from distinct sources of data are entered into the regression distinctly. (Zhang, Iwashyna and Christakis 1999) Further, we found that these improvements in explanatory power were largely true whether or not one also controlled for Charlson scores based on self-reported health history and / or based on the secondary diagnoses from the claim for the index hospitalization. Therefore we computed separate Charlson scores for each data source for 1-year intervals prior to each index admission date – this means that for individuals empanelled with multiple diseases, they have multiple, diagnosis-specific Charlson scores. Thus for an individual hospitalized on July 1, 1993, with an M.I., we have computed three hospitalization-claims-based Charlson scores: 1 for the year July 1, 1992 – July 1, 1993, a second for the year July 1, 1991 – July 1, 1992, and a third for the year July 1, 1990 – July 1,1991; we have computed parallel scores in other claims types.

A.4 Characterizing the Initial Hospitalization

Table A.7 presents some basic information on the initial hospitalizations of the COSI cohort members. Data is presented for the temporally first hospitalization for those with multiple empanelling diseases; it is also restricted to only those probands who lived in the 50 states or D.C. As is clear, there is substantial

heterogeneity both in terms of median lengths of stay and the variability within diseases in length of stay. Patients with urinary tract cancers (which exclude prostate cancers) had the shortest median length of stay; colorectal cancer patients had the longest stay. There was also substantial heterogeneity in the end-points of the initial hospitalization. Many M.I. and C.N.S. cancer patients were transferred to other acute inpatient hospitals; quite few of the other cancer patients were so transferred. Patients with M.I., stroke, lymphoma or malignancies of the liver and biliary tract, lung, or pancreas all had less than a 1 in 10 chance of surviving their initial hospitalization. In contrast, patients with urinary tract cancer and hip fracture had better than a 1 in 20 chance of surviving. There were low levels of disagreement between the claims and the vital status records as to whether or not a patient died at discharge; approximately 0.5% - 1.5% of the claims stated that the patient was "discharged to death" when the vital status records indicated the patient died at least 2 days away from the discharge date. Our policy was to trust the vital status records, as these were used for administrative purposes (such as Social Security eligibility) while that "discharge destination" field of the claims is not, to the best of our knowledge, used for reimbursement. Linkage to the National Death Index would be needed to provide an empirical foundation for this position.

A.4.1 External Data Linkages: Hospitals

There were a total of 5,103 hospitals in the MedPAR data; not all of these hospitals were included in COSI. 5,084 had at least 10 Medicare discharges in 1993

and could be identified in HCFA's Provider of Service File. From this information, we linked to the 1993 American Hospital Association (AHA) Annual Survey data. (American Hospital Association 1994) The AHA data is a survey of all hospitals; it is typically considered the best self-reported source of information on hospital features. Using hospital names, local address, and telephone number, from HCFA, we were able to link to a total of 4,923 (96.8%) short-term acute care hospitals in the AHA annual survey database.

A.4.2 External Data Linkages: Individual Patients

A major limitation of claims-based data explorations is the paucity of individual-level information about non-health-related attributes or outcomes. In the current project, we have attempted to overcome this in two ways. First, we have tried to maximally exploit the information available from HCFA, using the marriage detection algorithm, expanded ethnicity detection algorithms, and detail comorbidity measures. Second, we have taken advantage of the many high-quality local area data sets available from the U.S. Government: in particular, we link to the 1990 Decennial Census and the Area Resource File. While this auxiliary data sets do not provide individual-level detail, they provide important information about the communities in which our probands make their lives. For many studies, this local area information is quite useful.

A.4.2.1 U.S. Census

Data were linked to the 1990 U.S. Decennial Census; the Census provides the most detailed information about population characteristics available. (U.S. Dept. of Commerce Bureau of the Census 1991) This was done at the ZIP-code level. ZIP-codes are aggregations of 25,000 to 50,000 residents developed for administrative purposes. As such, they do not necessarily represent community boundaries, in the way community areas or census tracts attempt to. However, because of their ready availability and relatively low level of aggregation, they are often used in linkage studies. (Alexander and Sehgal 1998; Carlisle and Leake 1998; Feinglass et al. 2000; Garcia et al. 2000; Kaestner, Racine and Joyce 2000; Philbin et al. 2000; Roetzheim et al. 1999) We were able to link 1,184,995 (97.1%) of the 1,221,153 probands who were in the 50 states and D.C. to the 1990 Census. The linkage failures likely result from data errors in the claims and the Post Office's periodic creation of new ZIP codes in dense areas. From this, we were able to extract information about the communities in which the probands reside, such as the age distribution, race, median income, median education-level, and population density.

We were particularly interested in the use of Census data to provide additional information on the level of affluence of the communities in which our probands reside. This provides a continuous measure that is likely well-correlated with household-level total financial resources. The interpretive validity of this approach has been validated; (Hofer et al. 1998; Krieger 1992) however, there remain certain limitations as to the interpretation of any estimated effects from such proxy values.

(Geronimus and Bound 1998; Geronimus, Bound and Neidert 1996; Robinson 1950) There is an extensive debate on the usefulness of such area-based measures in the literature. (Davey Smith, Ben-Shlomo and Hart 1999; Greenwald et al. 1994; Hyndman et al. 1995; Krieger and Gordon 1999; Summer and Wolfe 1978) The major interpretive difficulty comes because geographical data may tend to under-control for variation in economic resources – for example, it will fail to take into account the fact that African-Americans in general have lower levels of wealth at the same income levels as whites. (Oliver and Shapiro 1995) However, among the elderly, the use of area-measures may better approximate the concept of mobilizable financial resources – such a home equity – than would a simple measure of income. For health decisions, particularly at the end-of-life, a more general measure of assets may be more appropriate for studying the influence of finances on choices.

A.4.2.2 Area Resource File: County-level Definition of Market Variables

The Area Resource File is a publicly available aggregation of data from a number of sources produced by the federal government. It is commonly used in health services research to provide information at the county level. (Banaszak-Holl, Zinn and Mor 1996; Halfon et al. 1996; Hartley, Moscovice and Christianson 1996; Kerstein, Pauly and Hillman 1994; Lafata, Koch and Weissert 1994; Lambrew and Ricketts 1993; Mullan, Politzer and Davis 1995; Roetzheim et al. 1999; Succi, Lee and Alexander 1997; Wholey et al. 1997) We were able to link 1,203,919 (98.6%) of the 1,221,153 probands who resided in the 50 states and the District of Columbia.

There was no particular relationship between whether patients could be linked to the Census via ZIP-codes or linked to the A.R.F. via county information. From the Area Resource File we could obtain a number of health care infrastructure, population demographics, and other variables

Counties were particularly of interest to us as we wanted to define the health care markets in Chapter 4. There are a number of difficult methodologic issues involved in defining health care markets. Some have strongly advocated the use of the Hospital Referral Regions, (Wennberg and Cooper 1998) others the use of network-based measures, (Phibbs and Robinson 1993; Sohn 1996; Succi, Lee and Alexander 1997) and others counties. In this project we have used counties to approximate markets, as has been done in numerous other studies. (Banaszak-Holl, Zinn and Mor 1996; Halfon et al. 1996; Hartley, Moscovice and Christianson 1996; Kerstein, Pauly and Hillman 1994; Lafata, Koch and Weissert 1994; Lambrew and Ricketts 1993; Mullan, Politzer and Davis 1995; Murtaugh 1994; Padgett et al. 1994; Roetzheim et al. 1999; Succi, Lee and Alexander 1997; Wholey et al. 1997) This was done for a number of reasons: (1) our experience with patients suggests that counties best approximate the way they think about their market's boundaries; (2) empirical tractability and availability of data; and (3) past work suggesting that results are often (but not always) insensitive to the difference between HRRs and counties. (McLaughlin et al. 1989)

Figure A.1: Overview of Data Construction

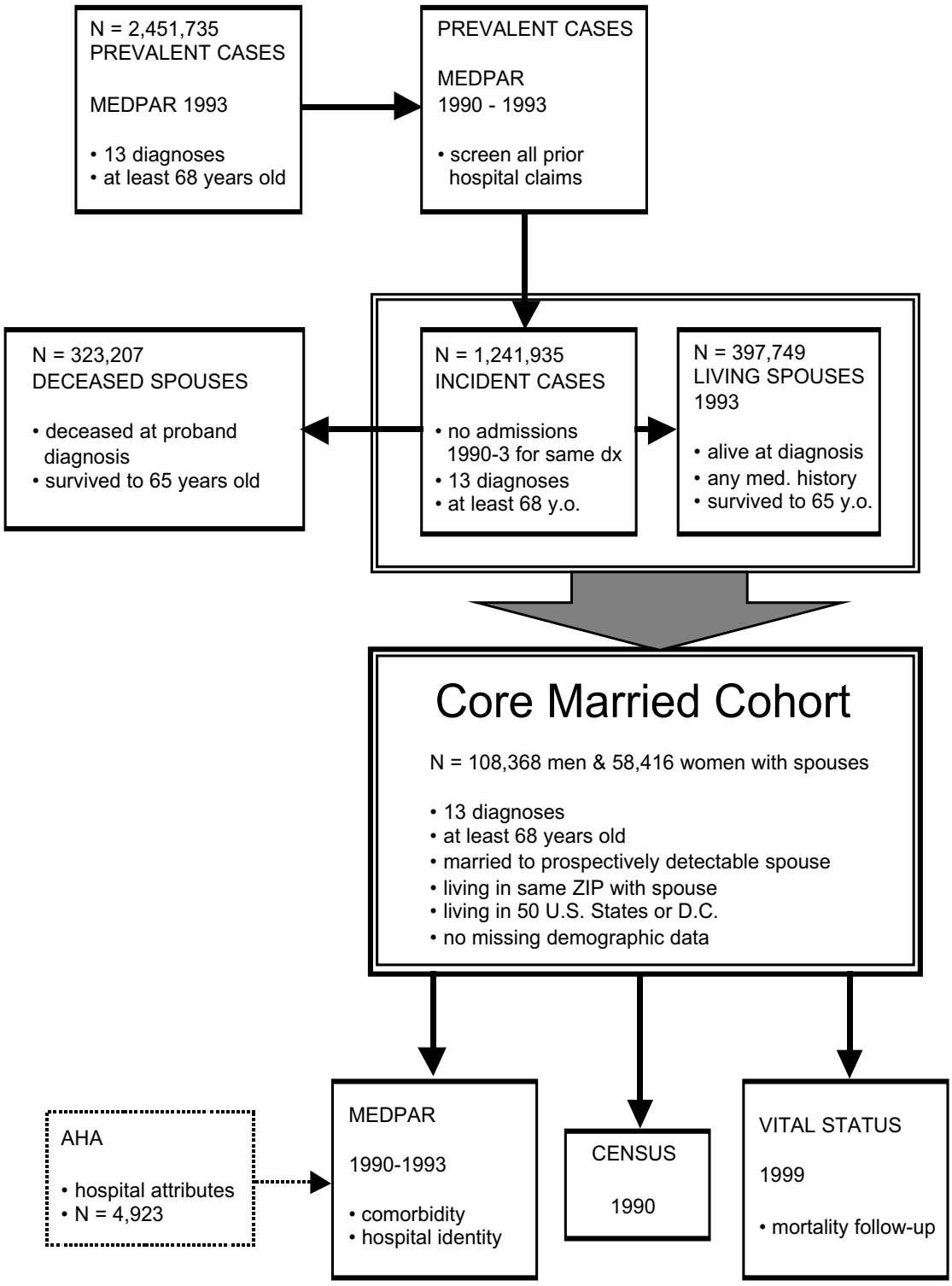

Table A.1: ICD-9-CM Operationalizations of COSI Diagnoses

	ICD-9 Codes	Source	Others Considered
Non-Cancer			
Heart Attack	410.0-410.9 (exclude 410.x2)	Jollis et al. 1996; Krumholz et al. 1998	Asch et al. 1995; Draper et al. 1990; Fisher et al. 1992
C.H.F.	398.91,402.01, 402.11,402.91, 404.01,404.03, 404.11,404.13, 404.91,404.93, 428.0-428.9	Taylor Jr., Whellan and Sloan 1999	Asch et al. 1995; Draper et al. 1990; Elixhauser et al. 1998; Fisher et al. 1992; Krumholz et al. 1997
Hip fracture	820-820.9	Fisher et al. 1992	Lauderdale et al. 1997
Stroke	434, 436	Benesch et al. 1997; ** Holloway et al. 1996	Asch et al. 1995; Draper et al. 1990; Fisher et al. 1992; Lee, Huber and Stason 1996; Taylor Jr., Whellan and Sloan 1999; Wolinsky et al. 1998
Cancer			
Colorectal	153-154.8	Fisher et al. 1992	McBean, Warren and Babish (1994) use colon only
Lung	162.2-162.9	McBean, Babish and Warren 1993, 1994	
Urinary *	188-189	Ries et al. 1997	
CNS	191, 192, 194.3, 194.4	our own	Counsell, Collie and Grant (1997), Loomis and Savitz (1990) and McKinney et al. (1994) used intracranial tumors only
Head/Neck	140-149, 161	our own	Allison, Franco and Feine (1998), Allison, Locker and Feine (1998), and Ries et al. (1997) were combined for this project.
Leukemia	204-208.9	Loomis and Savitz 1990; Ries et al. 1997; Ventrees and Manton 1986	
Lymphoma	200-203; 238.6	Ries et al. 1997	
Liver/biliary	155-156	Ries et al. 1997	
Pancreatic	157	Ries et al. 1997	

* Urinary Tract cancers do *not* include prostate cancers.

** Benesch et al. (1997) provided a comparison of multiple definitions with chart-review.

Sources:

Allison, P. , E. Franco, and J. Feine. 1998. "Predictors of professional diagnostic delays for upper aerodigestive tract carcinoma." *Oral Oncolog* 34:127-132.

Allison, P., D. Locker, and J. S. Feine. 1998. "The role of diagnostic delays in the prognosis of oral cancer: a review of the literature." *Oral Oncology* 34:161-170.

Asch, Steven, Elizabeth Sloss, Richard Kravitz, Caren Kamberg, Barbara Genovese, and Roy Young. 1995. "Access to Care for the Elderly Project.". Santa Monica: RAND.

Benesch, C, D.M. Witter Jr., A.L. Wilder, P.W. Duncan, G.P. Samsa, and D.B. Matchar. 1997. "Inaccuracy of the International Classification of Disease (ICD-9-CM) in identifiying the diagnosis of ischemic cerebrovascular disease." *Neurology* 49:660-664.

Counsell, Carl E., Donald A. Collie, and Robert Grant. 1997. "Limitations of Using a Cancer Registry to Identify Incident Primary Intracranial Tumors." *Journal of Neurology, Neurosurgery, and Psychiatry* 63:94-97.

Draper, David, Katherine L. Kahn, Ellen J. Reinisch, Marjorie J. Sherwood, Maureen F. Carney, Jacqueline Kosecoff, Emmett B. Keeler, William H. Rogers, Harry Savitt, Harris Akken, Kenneth B. Wells, David Reboussin, and Robert H. Brook. 1990. "Studying the Effects of the DRG-Based Prospective Payment System on Quality of Care: Design, Sampling and Fieldwork." *JAMA* 264:1956-1961.

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36:8-27.

Fisher, E.S., F.S. Whaley, M. Krushat, D.J. Malenka, C. Fleming, Baron J.A., and D.C. Hsia. 1992. "The Accuracy of Medicare's Hospital Claims Data: Progress Has Been Made, but Problems Remain." *American Journal of Public Health* 82:243-248.

Holloway, R. G., D. M. Witter Jr., K. B. Lawton, J. Lipscomb, and G. Samsa. 1996. "Inpatient costs of specific cerebrovascular events at five academic medical centers." *Neurology* 46:854-860.

Jollis, James G., Elizabeth R. DeLong, Eric D. Peterson, Lawrence H. Muhlbaier, Donald F. Fortin, Robert M. Califf, and Daniel P. Mark. 1996. "Outcome of acute myocardial infarction according to the specialty of the admitting physician." *New England Journal of Medicine* 335:1880-1887.

Krumholz, Harlan M., Eugene M. Parent, Nora Tu, Viola Vaccarino, Yun Wang, Martha J. Radford, and John Hennen. 1997. "Readmission after Hospitalization for Congestive Heart Failure among Medicare Beneficiaries." *Archives of Internal Medicine* 157:99-104.

Krumholz, Harlan M., Martha J. Radford, Yun Wang, Jersey Chen, Asefeh Heiat, and Thomas A. Marciniak. 1998. "National Use and Effectiveness of the Beta-

Blockers for Treatment of Elderly Patients After Acute Myocardial Infarction." *JAMA* 280:623-629.

Lauderdale, Diane S., Steven J. Jacobsen, Sylvia E. Furner, Paul S. Levy, Jacob A. Brody, and Jack Goldberg. 1997. "Hip Fracture Incidence among Elderly Asian-American Populations." *American Journal of Epidemiology* 146:502-9.

Lee, A. James, Joyce Huber, and William B. Stason. 1996. "Poststroke Rehabilitation in Older Americans: The Medicare Experience." *Medical Care* 34:811-825.

Loomis, Dana P., and D. A. Savitz. 1990. "Mortality from brain cancer and leukaemia among electrical workers." *British Journal of Industrial Medicine* 47:633-638.

McBean, A. Marshall, J. Daniel Babish, and Joan L. Warren. 1993. "Determination of Lung Cancer Incidence in the Elderly using Medicare Claims Data." *American Journal of Epidemiology* 137:226-234.

McBean, A. Marshall, Joan L. Warren, and J. Daniel Babish. 1994. "Measuring the Incidence of Cancer in Elderly Americans Using Medicare Claims Data." *Cancer* 73:2417-2425.

McKinney, P. A., J. W. Ironside, E. F. Harkness, J. C. Arango, D. Doyle, and R. J. Black. 1994. "Registration quality and descriptive epidemiology of childhood brain tumours in Scotland 1975-90." *British Journal of Cancer* 70:973-979.

Ries, Lynn A. Gloeckler, Carol L. Kosary, Benjamin F. Hankey, Barry A. Miller, Angela Harras, and Brenda K. Edwards (Eds.). 1997. *SEER Cancer Statistics Review, 1973-1994*. Bethesda, MD: National Cancer Institute. NIH Pub. No. 97-2789.

Taylor Jr., Donald H., David J. Whellan, and Frank A. Sloan. 1999. "Effects of Admission to a Teaching Hospital on the Cost and Quality of Care for Medicare Beneficiaries." *New England Journal of Medicine* 340:293-9.

Ventrees, J., and K.G. Manton. 1986. "The complexity of chronic idsease at later stages: practical implications for prospective payment and data collection." *Inquiry* 23:154-165.

Wolinsky, F. D., G. J. Wan, J. G. Gurney, and D. W. Bentley. 1998. "The Risk of Hospitalization for Ischemic Stroke Among Older Adults." *Medical Care* 36:449-461.

Table A.2: Alternative Definitions of Incidence: Unique Index Hospitalizations

	1° position		any position	
	prevalent	incident	prevalent	incident
Non-Cancer				
Heart Attack	**256,183**	234,098	323,736	296,144
CHF	494,845	299,161	1,294,707	833,027 (sample = **277,676**)
Hip fracture	218,729	207,927	**228,677**	216,431
Stroke	**268,222**	241,479	334,016	300,093
Cancer				
Colorectal	78,189	72,165	98,877	**84,093**
Lung	58,077	51,072	110,243	**87,619**
Urinary Tract	39,553	31,142	54,964	**40,897**
"Bad" Cancers:				
Leukemia	9,505	7,168	34,940	**22,017**
Lymphoma	22,182	16,671	53,042	**34,327**
Pancreatic	12,834	11,661	19,233	**16,225**
Liver/biliary	7,383	6,695	11,290	**9,655**
CNS	5,636	5,103	7,230	**6,276**
Head/Neck	9,051	7,848	14,127	**11,428**

Bold numbers indicate the choice made for this project. This is based on acute inpatient hospitalizations with age at least 68 for any hospitalization (that is, without any geographic restrictions, claims completeness, or date validity checks)

Table A.3: Alternative Estimates of the Incidence of COSI Diagnoses in the Elderly, in thousands of events per year

	SEER	1993 MedPAR *	COSI
Colorectal	91	95	84
Lung	125	104	86
Leukemia	16	24.5	22
Lymphoma	38	39	34
Pancreatic	20	18	16
Liver/biliary	12.5	11.6	10
CNS	6.3	8.8	7.4
Head/Neck	14.8	14.3	11.6
Urinary Tract	59	47	41

* 1993 MedPAR, *any age*, no look-back to exclude incident cases.
Source: SEER data is from Ries, Lynn A. Gloeckler, Carol L. Kosary, Benjamin F. Hankey, Barry A. Miller, Angela Harras, and Brenda K. Edwards (Eds.). 1997. *SEER Cancer Statistics Review, 1973-1994*. Bethesda, MD: National Cancer Institute. NIH Pub. No. 97-2789. Other columns are from authors' own tabulations.

Table A.4: Number of Index Hospitalizations in 1993, by Unique Individuals

Count	Frequency	Percent of Individuals
1	1,144,365	92.1 %
2	93,892	7.6 %
3	3,568	0.3 %
4	109	0.0 %
5	0	0.0 %
6	1	0.0 %

This is based on acute inpatient hospitalizations with age at least 68 for any hospitalization (that is, without any geographic restrictions, claims completeness, or date validity checks)

Table A.5: Marital Status of All COSI-cohort as of Jan. 1, 1993

Men			Who Died First?		
	Total	% of Men	Proband	Spouse	Neither
Method I					
Proband is A	124,217	23.9%	24,334	73,376	26,507
Proband is B	3,304	0.6%	589	1,966	749
Method II					
Proband is A	119,299	22.9%	22,161	89,094	8,044
Proband is B	3,427	0.7%	942	2,064	421
Follow-Up					
Proband is A	28,021	5.4%	1,500	18,595	7,926
Proband is B	263	0.1%	23	141	99
Widowed	54,242	10.4%			
Unmatched	187,049	36.0%			
Total	519,822				

Women			Who Died First?		
	Total	% of Women	Proband	Spouse	Neither
Method I					
Proband is A	1,907	0.3%	641	839	427
Proband is B	68,567	9.6%	25,195	29,410	13,962
Method II					
Proband is A	5,387	0.8%	2,248	2,936	203
Proband is B	40,399	5.7%	16,942	17,959	5,498
Follow-Up					
Proband is A	654	0.1%	83	411	160
Proband is B	2,304	0.3%	833	983	488
Widowed	268,965	37.8%			
Unmatched	323,889	45.5%			
Total	712,072				

Table A.6: Death by End of 1997 Among Decedents

	Of Probands, Who Died by Dec. 31, 1997?	
	Men	Women
Heart Attack	53.9%	57.0%
C.H.F.	71.9%	65.8%
Hip Fracture	70.8%	54.6%
Stroke	64.2%	62.6%
Cancers		
CNS	95.1%	90.6%
Colon	61.6%	57.6%
Head & Neck	72.7%	67.8%
Liver & Biliary Tract	94.6%	93.0%
Leukemia	81.4%	75.9%
Lung	92.4%	87.6%
Lymphoma	80.4%	75.7%
Pancreas	96.4%	95.0%
Urinary Tract	57.7%	58.4%
Overall	68.0%	62.6%
Number of Cases	514,732	706,421

This requires that the probands lived in the 50 states or D.C. Probands are tabulated only once, by their temporally first diagnosis if they were multiply empanelled.

Table A.7: Characteristics of Initial Hospitalization

		Length of Stay (days)			End of Stay		
	N	Median	25%	75%	Transfer	Death (V.S.)	Death (Claims)
Heart Attack	218,946	7	4	10	15.3%	15.8%	16.5%
C.H.F.	253,093	7	4	11	3.8%	10.1%	10.6%
Hip Fracture	210,493	8	6	11	5.6%	4.0%	4.2%
Stroke	244,259	7	4	10	5.9%	9.7%	10.1%
Cancers							
CNS	5,536	8	4	14	8.2%	7.7%	8.0%
Colon	80,209	10	7	15	1.2%	6.3%	6.6%
Head & Neck	10,565	6	3	12	1.7%	5.9%	6.1%
Liver & Biliary Tract	8,504	9	5	15	3.7%	15.7%	16.4%
Leukemia	20,489	6	4	11	3.9%	14.9%	15.5%
Lung	83,888	8	4	13	2.4%	13.4%	13.7%
Lymphoma	31,630	7	4	13	2.8%	9.0%	9.5%
Pancreas	14,993	9	5	16	2.9%	15.7%	16.1%
Urinary Tract	38,548	5	3	9	1.2%	3.4%	3.5%

Death (V.S.): Proband's death date as recorded in the Vital Status file is the same as discharge date from claims.
Death (Claims): Claim recorded proband's discharge disposition as discharged to dead.