Comment   Ernst R. Berndt

Chapter 8 in this volume addresses a very important set of issues regarding implications of using retrospective medical claim transactions data to construct price indexes for the treatment of episodes of various randomly chosen medical conditions and diseases. Price index data initially collected by the Bureau of Labor Statistics (BLS) in Boston, New York City, and Philadelphia for use in its medical Consumer Price Index (CPI) are compared with medical claims data in the Thomson Reuters MarketScan Research Databases, the latter based on transactions with providers/retailers in the same three cities, all over the January 1999 to December 2002 time period. That the BLS commission this type of experimental study was recommended by the panel convened by the Committee on National Statistics.[1] This chapter represents the first empirical evidence on the feasibility of constructing disease treatment-based price indexes that allow for substitutability among medical inputs and on comparisons of their price trends with those published by the BLS in its medical CPI.[2]

The authors construct a number of alternative price indexes, based on forty conditions randomly selected using expenditure versus population weights (the latter a simple count of the number of episodes treated), small samples (the same size as those used by the BLS in collecting data for the CPI), and large samples (about ten times the BLS sample size). Using medical CPI standard error data provided by the BLS and standard error estimates for the episode-based price indexes obtained by implementing bootstrapping procedures, the authors then test whether any of the cumulative forty-eight-month price trends are statistically significantly different.

The central finding reported is that for the most part, while there appear to be very different trends among cities and methods over shorter time periods, after forty-eight months, the cumulative estimated price changes for the various methods are typically not statistically significantly different. In

1. Schultze, C., and C. Mackie, eds. 2002. *At what price? Conceptualizing and measuring cost-of-living and price indexes.* Washington, DC: National Academies Press.

2. I note in passing that the version of the paper given at the Vancouver conference had a slightly different title and contained considerably different material than was in the initial revised version from late 2006. The version I comment on here is dated July 1, 2007, and it already incorporates responses to some of my comments made in 2006. The Vancouver version was a draft by Xue Song, William D. Marder, Onur Baser, Robert Houchens, Jonathan E. Conclin, and Ralph Bradley, entitled "Can Health Care Claims Data Improve the Estimation of the Medical CPI?" dated June 17, 2004.

general and with several exceptions, while point estimates of the disease-based price indexes tend to suggest smaller price index growth after forty-eight months than does the medical CPI for each of the three cities, the boostrap-based estimated standard errors are very large, resulting in the inability to reject the null hypothesis of no difference between them at usual p-values. Specifically, as reported in table 8.4, while the BLS all-medical-item CPIs have cumulative increases of 38 percent in Philadelphia, 23 percent in Boston, and 7 percent in New York City, the expenditure-based disease indexes all have negative cumulative price changes: –4 percent in Philadelphia, –8 percent in Boston, and –10 percent in New York City. Although the differences between the two are extremely large, being –42 percent for Philadelphia, –31 percent for Boston, and –17 percent for New York City, only that for Boston (which has the largest claims sample size) is statistically significantly different from zero.

What is one to make of this initial evidence? I am reminded of my first statistics course professor, who exhorted us students never to forget that absence of evidence is not the same as evidence of absence. I say this for several reasons.

First, relatively little information is given on how the bootstrap method was implemented and on the validity of the assumptions on which it is based. Specifically, it appears that the authors' computation of bootstrap standard errors is based on the assumption of a random walk specification, which is well-known to generate nonstationarity. It would have been useful for the authors to test for nonstationarity using available unit root tests and to assess the robustness of their estimated bootstrap standard errors under alternative specifications, such as stationary ones around a trend. If the random walk hypothesis is empirically invalid, it likely results in exaggerated standard error estimates. Whether alternative specifications of the underlying stochastic processes would have resulted in sufficiently smaller standard errors to change qualitatively the general findings of no difference in cumulative price growth over time is of course unknown. Admittedly, the random walk hypothesis is commonly used in financial analyses of stock market movements, but its applicability to health care price changes is unclear. Careful and detailed discussion of this bootstrap specification and computational procedure would have been useful. While analysts of claims expenditure (not price) data have long noted that large outlier observations are a trademark in health care, the standard errors of price ratios reported here are cause for concern.

Second, the research initiative reported by the authors seems remarkably preliminary and incomplete. Why, for instance, is the statistical comparison undertaken only after forty-eight months, and not, for example, at yearly intervals? How would results have looked if the sample set of diseases were fixed for the four years, rather than being updated annually, or after every two years? How important is the abrupt change of weights from Decem-

ber of one year to January of the next year, when the basket of treatment inputs is updated for each disease? What would have been the consequences of smoothing the weights over that year end to the beginning of the next year time interval or of using some other overlapping methods? Why do the expenditure weights for inpatient, outpatient, and prescription drugs differ so much in both the CPI and claims data across cities and between the CPI and claims data for the same cities? How important are the various implementation methods that increase churning in contributing to the relatively large standard errors? If one formed a three-city aggregate of both the CPI and the various disease-based measures, would the price trends have become statistically significantly different, due to smaller standard errors? Although regional variations are of interest, in typical policy discussions, we are most concerned about measuring medical inflation at an aggregate rather than city-specific level of aggregation. Unfortunately, these types of rather obvious questions are not addressed in detail by the authors, and thus the reader is left rather puzzled, with the study raising enormous issues about both the CPI and Thomson Reuters underlying databases. Admittedly, the authors suggest that in future research, it might be useful to implement a two-stage disease selection method, first by major body organ system and then within that (an eminently reasonable suggestion that should not have been that difficult to carry out). In general, the analysis undertaken and reported in the chapter is spartan and truly preliminary.

A number of other issues deserve attention. First, while economic theory provides a strong rationale for using expenditure weights, what is the rationale for sampling based on population weights? The BLS and Bureau of Economic Analysis do not generally publish democratic weights (expenditure weights underlie plutocratic weighting schemes), so why even consider them here? Given the skewness of health care spending, one should expect substantial differences between expenditure- and population-weighted price indexes, and that is exactly what is reported. Why not instead devote more resources to a deeper analysis of the expenditure-weighted indexes? In any case, the expenditure-based and population-based indexes often display opposite trends for which no explanation is given.

Second, the careful reader will note from tables 8.1 and 8.2 that for the BLS medical CPI, in both Philadelphia (even-numbered months) and Boston (odd-numbered months), but not in New York City, the BLS only samples bimonthly. This creates a number of statistical complications when comparing their cumulative growth to all monthly disease-based measures and raises the issue of why the periodicity for the disease-based measures for Philadelphia and Boston differ from those of the BLS for those cities. A related puzzling result in table 8.3 is, why is it that the expenditure-based price index for prescription drugs goes up by 97 percent in Philadelphia and by 10 percent in Boston but drops by 39 percent in New York City? This regional variation in price changes is most puzzling.

Third, because the transaction claims data represent the total payment to the provider (consumers' out-of-pocket payments, plus third-party payer's payments to the provider), in some sense, the price index computed here better resembles a producer price index by medical care providers than a CPI, particularly because the medical care CPI relies so very heavily on a reallocation of insurance premiums into insurance payouts and gross margins.[3] While the Producer Price Index scope is generally confined to domestically produced goods and services, whereas the CPI tracks price changes of goods and services, regardless of where they are produced, my understanding is that for health care goods and services, relatively few are produced abroad, and thus the claims-based price index is more like a producer rather than consumer price index. A more complete discussion of these various issues would have been most useful.

Finally, a critical component of the analysis reported here involves the use of the Medstat Episode Grouper, which groups distinct medical claims over time into an episode of care, which in turn provides the basis for pricing inputs. The construction of such data into episode groupers involves combining medical knowledge with insights from claims processors and statisticians and reflects a great deal of art. It is my understanding that there are a number of alternative episode groupers available commercially (and some in the public domain), and thus it would have been useful for the authors to devote some attention in the chapter to the existing literature that examines and compares the various episode groupers on criteria such as their internal and external validity.[4]

In summary, the topic addressed by this chapter is extremely important. That so little in-depth analysis was undertaken, however, is disappointing. While the desirability of further research is typically mentioned in the concluding section of empirical chapters, in this case, that need is truly great.

3. For further discussion, see Schultze and Mackie, 2002, *At What Price?*

4. A study by Ana Aizcorbe, Nicole Nestoriak, and others that compares price indexes from alternative episode groupers (including the Symmetry Episode Treatment Group product commercially available from Ingenix, a subsidiary of United Health Care; available at http://www .ingenix.com/content/File/EvolutionofSymmetry.pdf) is currently underway. Initial findings have been reported by Ana Aizcorbe and Nicole Nestoriak in a Powerpoint presentation at the Bureau of Economic Analysis Advisory Committee Meeting, entitled "Episode-Based Price Indexes: Plans and Progress," on May 4, 2007. Available at www.bea.gov.