NCHS Data Linkage Activities: Opportunities and Challenges for Population Health Research.

Kimberly A. Lochner and Christine S. Cox

Introduction

Federally sponsored health surveys are a critical source of information on population health in the United States. Yet, surveys must balance the range of health topics covered against the burden imposed on respondents. In addition, cost constraints often prohibit active follow-up of survey participants to assess their subsequent health status and mortality. For these reasons, linking individual survey records to other data sources provides a scientifically valuable and cost-effective means to enrich existing data collection activities. Recently the National Center for Health Statistics (NCHS) has expanded its data linkage program for its population-based surveys, such as the National Health Interview Survey (NHIS), by linking individual survey records to administrative data on mortality, Medicare enrollment and utilization, and Social Security benefit histories. These new data sources can advance our understanding of the multiple factors influencing population morbidity, disability, and mortality.

Although record linkage increases both the quality and quantity of data available, the ability to produce linked data sets that are scientifically valuable faces several challenges. For example, survey respondent identification data such as Social Security Number (SSN), full name, and date of birth is essential for accurate matching to administrative records. However increased public awareness of identity protection coupled with changes in NCHS survey administration protocols, with regard to informed consent and the collection of personal identification data, has resulted in significant increases in non-response to key identification data. The effect is fewer survey participants who are eligible for record linkage, which may result in biases in the linked data sample.

This paper will describe the new NCHS linked data files, including the eligibility criteria for linkage and the methodology used to match survey records to the different administrative resources. Also, this paper will describe the trend of increasing SSN refusal among NHIS respondents and linkage rates of NHIS respondents to mortality, Medicare, and Social Security records, including sub-group patterns in linkage rates for key socio-demographic and health characteristics.

Data and Methods

NCHS population based surveys, including the NHIS, the Longitudinal Study of Aging (LSOA II), and the National Health and Nutrition Examination Surveys (NHANES) are linked to administrative data from death certificates, Medicare, and Social Security. This paper focuses on the NHIS linked data files for several reasons. First, the NHIS is the principal source of information on the health of the civilian non-institutionalized population of the United States. Second, the NHIS is an annual survey and has 15 years of data linked to death information, which is particularly important for studies examining socioeconomic or race/ethnic differences in mortality and survival, as well as five years (1994-1998) linked to Medicare and Social Security records, which allows for studies on

functioning, well-being, and disability among the elderly. Finally, linking NHIS records highlights the challenges of producing valid linked data samples as the percentage of NHIS respondents who refuse to provide or are missing on key identification data used for linkage have significantly increased in recent years. All analyses take into account the complex sample design and weights using the SUDAAN software application.

Findings

The NCHS data linkage activities have created several new linked data files that augment the available information for demographic research by increasing the accuracy and detail of the data collected and providing a longitudinal component to NCHS health surveys. These data resources are important for studies of trends in the socioeconomic gradient in health, racial differences in mortality, and trends in well-being and functioning in the elderly.

The growing challenges to linking the NHIS to mortality and other data sources is evident by the decreasing proportion of respondents who are eligible for linking to other data sources. Survey participants were not eligible for linking to administrative data, if at the time of their interview, they refused to provide their Social Security Number (SSN) or if they refused, were missing, or had incomplete information on last name and date of birth. The proportion of NHIS respondents who provide their SSN has decreased dramatically over the last decade, from 70% in the early 1990's to less than 50% in 1998 to roughly 25% in 2003. The result is that the proportion of adults ineligible for linkage in the NHIS Linked Mortality files increased from less than 3% in 1995 and earlier to over 10% by 1998. A similar pattern is evident for the 1994 to 1998 NHIS respondents linked to Social Security and Medicare data. Although among those eligible for linking, the match rate is above 90%, the total match rate drops from 75% in 1994 to 55% in 1998. The decrease is due almost entirely to the proportion ineligible for linkage increasing from 18% in 1994 to 40% in 1998.

In addition, the characteristics of survey respondents who refuse to provide personal identifying information differ systematically from those who do provide such information, which can lead to biases in the linked sample. For example, in the NHIS Linked Morality files, NHIS respondents who were younger, female, with higher socioeconomic status are less likely to be eligible for mortality follow-up. Since these groups have lower mortality rates, without adjustments to the sample weights, overestimation of mortality rates may occur. In the NHIS linkages to the Social Security and Medicare data, there were statistically significant differences in the proportion linked by key sociodemographic and health characteristics. For example, in the 1994 NHIS linked to Social Security data, 76% of non-Hispanic whites were linked compared to 69% of non-Hispanic blacks as were 78% of those with family incomes less than $10,000 compared to 72% for those with incomes greater than $50,000 – respondents who have "unknown" income have the lowest linkage rates (50%). Also respondents without health insurance were less likely to be in the linked sample and those with poorer health status were more likely to be linked.

Conclusion

The data linkage activities of NCHS have resulted in several comprehensive population-based systems of linked health records that provide a research rich environment to examine the behavioural, social and economic factors influencing population morbidity, disability, and mortality. However, potential biases may arise in linked data from survey respondents becoming less willing to provide personal identification data necessary for eligibility and accurate matches. Ways to address the challenges in producing scientifically valid linked data include developing new sample weights to account for differential selection into eligibility status as well as improved survey administration to collect identification data.