Union Army Veterans, All Grown Up

Sven E. Wilson, Noelle Yetter, Christopher Roudiez, Heather DeSomer, Eric Hanss

Abstract

This paper overviews the demographic research possibilities made possible from a major NIA-funded program project, *Early Indicators, Intergenerational Processes, and Aging.* Data collection began over two decades ago on a core random sample of almost 40,000 soldiers from the Union Army in the US Civil War. The collection is unique because it contains extensive demographic, economic, and medical data on these soldiers from childhood to death. In recent years that core longitudinal data has been expanded to include a large sample of African-American soldiers and an oversampling of soldiers from major US cities. Hundreds of historical maps containing public health data have been geocoded to allow placement of soldiers and their family members in a geospatial context. With newly granted funding, the project moves to the next generation as thousands of veterans will be linked to the demographic information available from the census and vital records of their children.

1. Overview

Only a data source going far back in time can give a complete life course perspective on the human population. For over two decades, researchers associated with the NIA-funded *Early Indicators* project have been accumulating and digitizing data on veterans of the Union Army (UA). The UA veterans, first identified as children, are observed in all the federal census schedules from 1850-1930, generate extensive data during their wartime service, and later become part of a federal pension system that collects detailed health, demographic and economic data until they die. Supplemental collections have augmented the original sample of almost 40,000 white recruits with over 21,000 African-American soldiers and an oversampling of 12,000 recruits from major cities in 19th century America.

In recent years, the UA sample has been complemented with a vast ecological dataset that contains a century's worth of public health, economic, and demographic detail from large US cities. Researchers have merged thousands of city maps, including detailed street-level changes over time, with thousands of tables of data that are linked to specific urban locations or to wards and neighborhoods. Because urban veterans in the pension system have street addresses in their pension files, veterans can be precisely located and geocoded. This allows analysis not only of urban living patterns but also of the environmental risk factors faced by urban veterans.

Over the next five years, the research team will use powerful new genealogical tools to link tens of thousands of children of UA veterans to their death records and to the federal census from 1860-1940, as well as constructing a similar non-UA intergenerational sample from published family histories. These new efforts allow a truly intergenerational and geospatial analysis of demographic processes.

This paper illustrates the wide-ranging research potential now available in the *Early Indicators* collection and highlights the exciting new developments and ongoing collections that will soon be available to researchers. For about three decades, hundreds of researchers have painstakingly created an unparalleled store of data that captures the complete lives of over 70,000 soldiers—black and white, urban and rural, immigrant and native-born—and built around those life histories a rich source of ecological data. Though many papers have been published using the *Early Indicators* collections, in many respects it is a collection that is only now coming of age.

2. Union Army Veterans and Related Collections

In the 1980s, a team of researchers led by the late Robert Fogel started exploring the feasibility of creating a national random sample designed to capture life course data on soldiers who served in the Union Army during the US Civil War. A grant application in 1986 was rejected on the grounds that such a massive and wide-ranging data collection project was not feasible. Fogel, ever the optimist, was determined to prove the critics wrong. With assistance of the National Bureau of Economic Research, he began collecting a pilot sample of 20 companies, developing at the same time the detailed software and collection methods necessary to capture the soldier's experiences during the war, their family background from the federal censuses, and their health histories found in the pension records housed in the National Archives.

A. Completed Veteran Collections

Armed with the evidence of a successful pilot, Fogel's 1991 application to the National Institute on Aging and the National Science Foundation, *Early Indicators of Later Work Levels, Disease and Death,* was successful. Since that time the original collection has been supplemented in important ways. Data from the United States Colored Troops (USCT) were added, and an oversample of urban recruits from the largest US cities was added to allow for a more detailed analysis of the effect of urban environments. POWs from the infamous Andersonville camp have also been included, allowing the study of acute stress. Finally, a collection of 1,700 veterans who lived to at least the age of 95 will soon be publicly available, which, along with long-lived veterans from the main samples, gives a new window on early determinants of extreme longevity. Each of these collections is described in brief in the paragraphs below.

1. Union Army Veterans (UAV)

The main thrust of that first major grant in 1991 was a sample of 39,340 recruits consisting of the complete enlistment of 332 randomly drawn companies in the Union Army. Recruits in the sample were enlisted white soldiers in infantry units from all Northern states. Wartime data are drawn first from the Compiled Military Service Records (CMSRs), which contain information on the recruit from his enlistment record (residence, occupation, age, height) and variables from a monthly muster record, which includes health, battles, wounds, hospitalization, desertion, POW status, and cause of death or muster-out. This data is supplemented by variables from the Carded Medical Records (CMRs), which contain information from military hospitals.

From the starting point of the CMSR and CMR records, researchers then go both forward and backwards in the soldier's life. Using the index of pensions, they locate the pension records for those soldiers who lived long enough to enter the pension. From the pension files and the accompanying "Surgeons' Certificates," researchers draw a wealth of information, often covering several decades, that includes demographic and economic information and the medical records that were conducted by the Pension Bureau to determine eligibility for pension support. Pension records can be dozens and sometimes hundreds of pages in length, and they constitute the primary source of data in the Union Army collection.

The final source of records are obtained from the federal census years. When the project began, available census years only went through 1910, so the veterans were linked forward to the 1900 and 1910 census and backwards to the 1850 and 1860 census records. These linkages allow demographic data to be collected not only for the veteran, but also his family. Newer samples of veterans have also been linked to the 1870 and 1880 census and the 1920 and 1930 censuses that have become publicly available since the collection project began.

2. United States Colored Troops (USCT)

The most important addition to the sample has been the collection of black veterans from the USCT. These were conducted in two waves, an initial sample of approximately 6,000 black soldiers and their white officers, which has recently been supplemented by an additional 15,000 black soldiers. African-American companies did not begin to be formed until mid-way through the war, and their wartime experiences were quite different from the white soldiers. However, the basic records and collection procedures for the black and white soldiers are essentially the same.

The legacy of slavery can be easily seen in many dimensions of the USCT sample. For instance, linkage rates to the 1850 and 1860 censuses were much lower, especially for recruits who were slaves prior to the war. In later life, black veterans faced discrimination in obtaining pension support, and their linkage rates to later census years were lower than for whites, partly because their life expectancy was lower than white veterans. The USCT collection is an invaluable source of data for studying the lives of black veterans in the decades following the war.

3. Urban Veterans Supplement

The Urban sample is a stand-alone oversampling of Civil War veterans who enlisted in the largest US cities. Drawn in proportion to city size in 1860, the sample was designed to allow researchers to examine intra-city disparities in environmental conditions and draw inferences about the impact of ward conditions on the recruits' life-cycle aging process. The five target cities are Boston, Chicago, Philadelphia, New York (including Brooklyn), and Baltimore. The Urban sample consists of over 12,500 soldiers, who were then located within these five cities at the ward level.

The varying quality of records available for each city necessitated different procedures for sample identification. Using Dyer's Compendium of the War of the Rebellion, State Adjutant Generals reports, and various online resources, researchers compiled a list of companies that had more than 50% of their recruits enlisting in each target city. Researchers then extracted names and identifying information from the Regimental Books for those companies until the sample size was reached. Military, pension, medical and census records were collected for these urban recruits using the same procedures as the main Union Army Veterans sample.

4. Andersonville POW Supplement

The CMSRs and CMRs provide information on stressors faced by soldiers in the army. We know the battles they fought in and have a record of their wounds and injuries. We also know the location and length of service. Regimental histories can also be exploited to gain more information on the wartime events encountered by the veterans of a given company.

Soldiers taken as POWs were under conditions of extreme stress, including, in some cases, long periods of malnutrition and even starvation. A recent supplement to the data includes the records of 1,000 survivors of the Andersonville POW Confederate prison. Andersonville was the most notorious Confederate prison and had a mortality rate of 40 percent. In 2007, our researchers began collecting data on survivors of Andersonville, using the index developed by the National Park Service.

The Andersonville sample is drawn from those recruits who survived to 1900. These survivors were linked to their CMSR, CMR, pension records and the eight federal census years between 1850 and 1930. This collection consists of about 1,000 veterans. Of these veterans 197 are linked to their siblings who were soldiers, whose complete data is also collected and will be included in a future data release.

5. Oldest-Old Supplement

Studying extreme longevity often requires drawing individuals from populations large enough to generate a sufficient number of very old members. Given 19th century life expectancy, even the large Union Army collections have a relatively small number of members reaching very old ages. The oldest-old is an over-sampling of 1,700 individuals with an age of death confirmed to be at least 95 years old.

To reach the target sample of 1,700 oldest-old, a list of nearly 6,000 potential nonagenarians was compiled from many sources, including gravestone databases, obituaries, newspaper accounts, veterans associations and the 1930 and 1940 censuses. Death dates were confirmed from the pension files and the veterans were linked to all the census years from 1850 to 1940.

The oldest-old supplement contains the same records present in the other Veteran's collections, but often additional records are available in later years, including physician affidavits and home visits by the Veterans Administration representatives. These records contain information on care arrangements, disabilities and cognitive limitations not generally present for younger veterans.

B. Historical Urban Ecological (HUE) Data

For more than a decade, the Early Indicators team has been collecting data on the public health environment of seven major US cities: Baltimore, Boston, Brooklyn, Chicago, Cincinnati, Manhattan, and Philadelphia, from 1830 through 1930. This collection of geospatial data was drawn from thousands of maps and published data tables and is now publicly available as the Historical Urban Ecological (HUE) data. Detailed changes in ward boundaries, street layouts and other built environment were meticulously traced over a century. No comparable street-level data for studying US urban history is available elsewhere.

Researchers selected the HUE study cities and variables in order to best analyze the effects of intra-urban health disparities and public health interventions on individual mortality and longevity as observed through the Union Army and USCT cohorts. The HUE data set includes ward boundary changes, street networks, and ward-level data on disease, mortality, crime, water, sanitation, commerce, industry, public works, property values and many other variables reported by municipal departments for each study city. Researchers scoured archives in each of the seven cities, as well as obtaining ward-level maps from the Library of Congress.

In sum, these materials constitute a framework that allows researchers to create accurate historical spatial and tabular data and perform geospatial and statistical analysis at many scales, giving researchers a deeper look into life in the rapidly changing cities of the nineteenth century. The HUE data by itself provides a fascinating and dynamic picture of urban American history. But even more exciting is the ability to link the Union Army collections to the HUE data. Veterans who resided in large cities and applied for the pension provided their street address, which allows us to precisely pinpoint the urban environment the veterans were living in.

C. New and Ongoing Collections

Sadly, Robert Fogel died in June of 2013, and leadership of the *Early Indicators* project passed to Dora Costa at UCLA. In 2014, the National Institute on Aging awarded a five year grant extension under a slightly different name*, Early Indicators, Intergenerational Processes, and Aging*. The change in the grant's name reflected the new push to create datasets appropriate for studying human development and aging within an intergenerational framework.

1. Veterans' Children's Census (VCC) Collection

The *Early Indicators* data provide a mechanism for studying intergenerational processes that is unavailable in modern data. We can study the aging and mortality not only of the veterans but also of their children. The VCC collects all possible data on veteran's children and spouses, including both census records and, where available, information from vital records.
The VCC consists of three new scientific projects, each of which collects data on children and spouses of veterans. The collection consists of three parts: 1) 1,882 POWs who survived to 1900; 2) 8,500 white non-POWs; 3) 4,500 African-American soldiers from the USCT. Each of these samples is constrained to veterans living until 1900. These collections come at considerable cost, since it takes, on average, about 5 hours of searching genealogical databases and inputting the results in order to collect the descendant data for a single Union Army veteran.

The aim of the VCC database is to facilitate the study of the causal mechanisms of intergenerational transmission of health, including the transmission of stressful events from parents to children. The data also create a valuable new object of study: the aging and longevity of women. In historical intergenerational databases, women are often neglected because it can be hard to track them once they marry and change their surnames. However, the initial findings from the VCC are that about 70% of veterans' daughters are linked to at least one census outside of their fathers' household, even when we do not have a married name from the pension record. Similarly, a high rate of linkage to death records is possible for the daughters of the veterans (though linkage to death certificates is a function of location and year of death).

2. Intergenerationally-Linked Aging Sample (ILAS)

Although the Union Army sample is broadly representative of the broader population because of the high rate of participation in the Army, it is still useful to utilize data that is not confined to veterans. The ILAS data were originally collected by economic historian Clayne Pope from printed family histories and genealogies from across the United States. The sample covers 39 family lineages consisting of about 15,000 households over time. The sample includes intergenerationally-linked records of 118,162 males and females born between 1577 and 1983. About 27% of the individuals in ILAS are currently linked to at least one of the federal censuses between 1850 and 1910.

The ILAS project will explore the feasibility of linking the sample individuals to the 1920-1940 censuses and recollecting the 1850-1910 census records, which were originally obtained without the assistance of modern online search capabilities. It will also merge ecological and macroeconomic variables available from the censuses and other sources. The ILAS is a valuable companion to the UAV because it contains information from the same cohort as the UAV as well as cohorts born before and after the Union Army cohort.

3. Research Possibilities

In addition to describing the multiple data sources and new research tools that have become available, this paper overviews a few of the exciting research questions that can be illuminated with the *Early Indicators* data collections, including social gradients in health, racial inequality, social network dynamics, mobility in time and space, the long-term economic, demographic and health effects of local environments and public health interventions, and the intergenerational transmission of health, well-being and longevity. We detail a few of the important demographic topics that can be addressed with the *Early Indicators* data.

3.A. Racial Inequality

The *Early Indicators* data sources provide a unique resource for studying racial disparities in health and mortality. A special collection of over 21,000 black soldiers from the US Colored Troops (USCT) allows direct comparison with white veterans on a number of health-related dimensions. The life expectancy of black and white soldiers provides an indicator of hardships suffered by the African Americans in the decades following the Civil War. Of soldiers surviving their wartime service, only 28.1% of black soldiers survive to appear in either the pension rolls or census records in 1900, compared to 44.1% of white soldiers. As an illustration of this disparity, Table 1 below shows 10-year mortality rates, by age, for white and black veterans alive in 1900 with a known age and death date (an expanded analysis would account for those censored cases where post-1900 data is available, but no death date exists).

Table 1: 10-year Mortality Rates by Race and Age, 1900-1910

Age in 1900	Whites (UA)	Blacks (USCT)
50-54	16.6%	26.5%
55-59	22.6%	33.2%
60-64	31.4%	40.0%
65-69	44.6%	53.4%
70-74	60.2%	67.5%

These basic results can be explored further with the demographic data from the 1900 and 1910 censuses and from the pension files, which also contain detailed medical examinations.

The *Early Indicators* data also contains detailed information from medical exams conducted to determine eligibility for and level of pension support at several points in time over the life course. Because the data from the USCT was recently expanded to include 21,000 soldiers, sample sizes are now sufficient to study racial differences in both common and uncommon conditions.

The pension files also provide a view on the social and institutional consequences of racial prejudice in 19th century America. Wilson (2010), for instance, studies the impact of informal liberalization in the pension system on black and white veterans. Long before Congress officially acted to liberalize pension eligibility in 1890, veterans' groups were exerting political influence to increase dramatically the numbers of people being covered by the system. But this informal liberalization applied primarily to whites. Costa (2010) finds that the responsiveness of retirement decisions and living arrangements to pension income was much higher for black veterans than for whites, indicating the harsh alternative black veterans faced without that income.

3.B. Socioeconomic Gradients

The data contain several indicators of socioeconomic status and background, including household wealth, literacy, occupation and place of residence. The health impacts of a life full of hard labor, for instance,_are poorly understood in today's highly-mechanized world. Occupations indicate important differences in socioeconomic background, and the occupations of Union Army soldiers can be tracked, in many cases, from their early life to death. Pension income also raised the economic welfare of many, particularly black veterans. By 1907, the pension becomes primarily aged-based, and it is possible to examine the effect of income on later-life mortality while controlling for the presence of chronic health conditions. Recent work (Salisbury 2014) has extended the analysis of income to widows of veterans.

3.C. Local Urban Environments

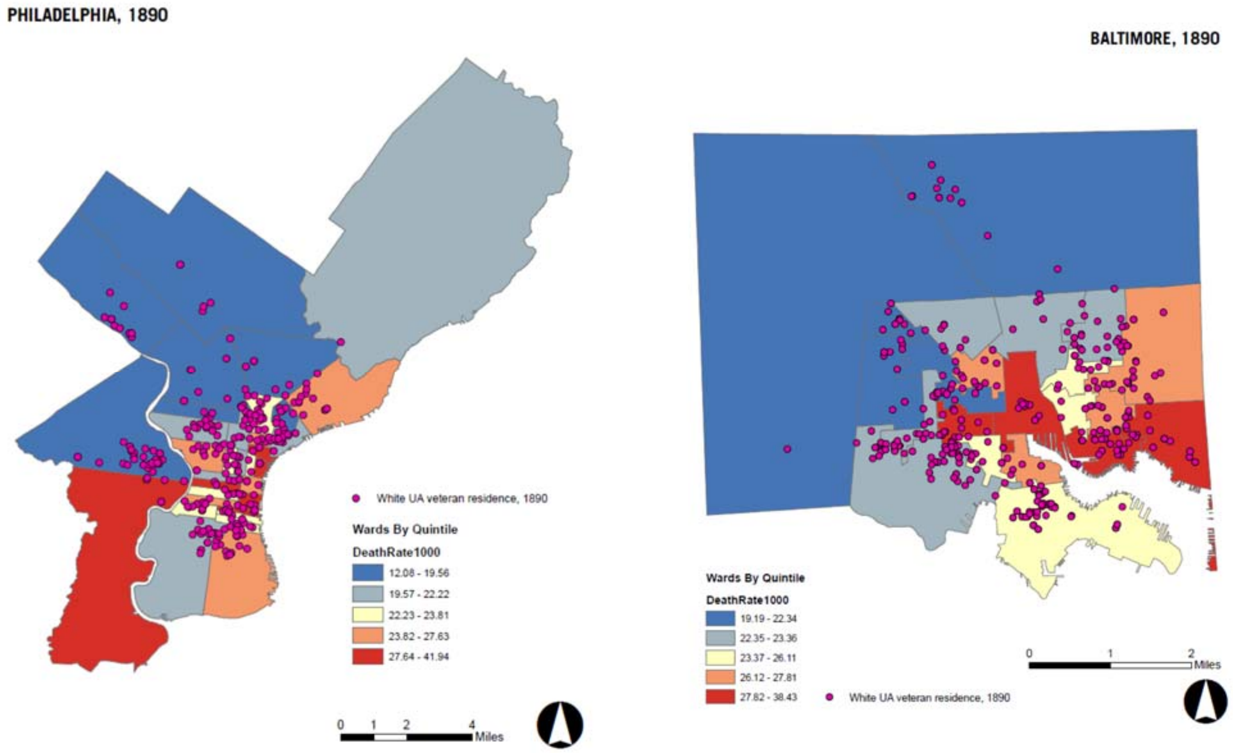
The *Early Indicators* collections provide an unprecedented perspective on the long-term effects of local environments on health and longevity. Recently expanded oversampling of Union Army recruits from major cities of the time allows for comparisons across cities and, more important, for analysis at the ward level. Researchers have spent many years geocoding historical maps so that recruits can be placed in a physical space that contains variables such as crime, disease, employment, government, municipal, property and vital. Many veterans return to their cities of origin, or they are part of the massive internal migration to the cities that took place in the late 19th century. Veterans, in many cases, can be mapped to the street-level and therefore linked to the geocoded public health information contained in the public health maps and the ward-level epidemiological and demographic data. Large-N studies that link individuals to their local neighborhood and public health data, such as infant mortality or infectious disease death rates, as well as tracking people over time, are not available from other historical or modern data sources.

A useful feature of studying these urban environments is that scholars can examine how social gradients affected health in an era where access to quality health care was not a confounding factor. Simply put, no one—rich or poor—had access to effective medical treatments for most conditions. Thus, the privileged and well-to-do, in contrast to today, faced many of the same health challenges as the poor. More privileged classes were able, however, to locate themselves in cleaner less congested environments. The *Early Indicators* collections allow for studying the impact of local environments at several points in time and space.

The HUE data allow the analysis of the local environment on health outcomes. The data contain a variety of health indicators, including specific diseases and overall mortality. As an example, the figure below shows the 1890 locations of veterans living in Philadelphia and

Baltimore in 1890. Colors on the map indicate the crude mortality rate of the ward, with red being the highest quintile and blue the lowest. 1890 is the year that many soldiers entered the pension system due to the liberalization of the law governing pension eligibility.

Figure 1: Veterans' Residences in 1890, by Ward Crude Mortality Quintile

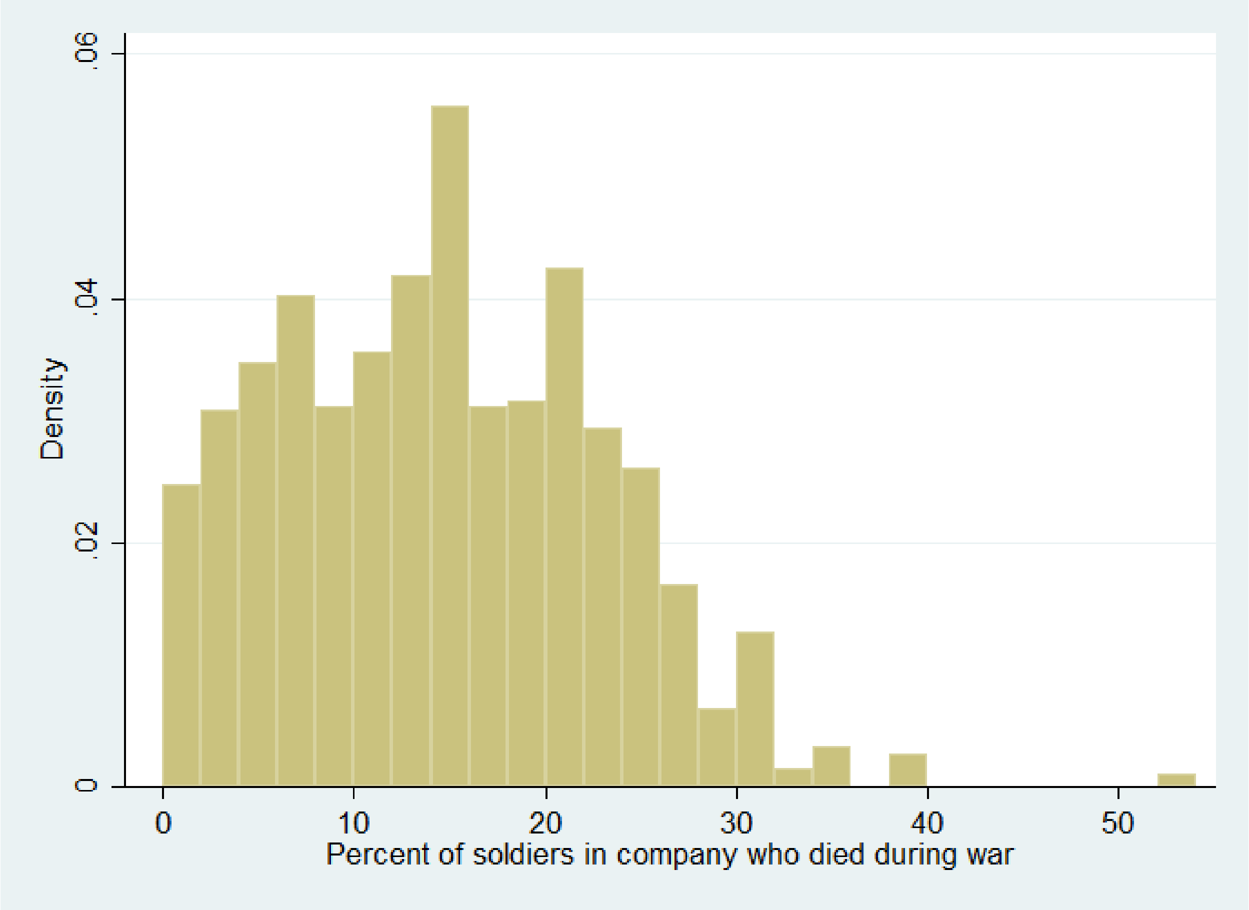


3.D. Acute Stress

A potential key indicator of later health and mortality is acute stress. Almost all soldiers in the Union Army experienced a measure of stress, but recruits varied significantly in their experiences. The military data files contain records of which battles were experienced by recruits as well as their POW history during the war.

A simple indicator of stress is the mortality rate experienced at the company level. The distribution of company-level mortality is shown in Figure 2 below, which indicates the broad range of experiences that companies experienced. This variation can be exploited as a measure of wartime stress. Researchers can also use information in the military files to construct measures of stress. These might include battles fought, locations in which the company served, injuries and wounds, and hospitalizations. Company-level mortality has been shown to have significant effects on rates of chronic disease (Pizzaro, Silver & Prause, 2006) and on later-life mortality (Costa & Kahn, 2010).

Figure 2: Company-Level Mortality during the Civil War



The intense stress of serving in POW camps can be studied with the Andersonville supplement discussed above. Costa and Kahn (2007) studied the survival of POWs in the camps finding that the presence of friends in the camp significantly increased the rate of survival. Costa (2012) has also examined the later life mortality of these same POWs. She finds that older age mortality is significantly linked to the age of imprisonment. Younger soldiers (age<30) face higher old-age mortality than their peers, but those imprisoned after age 30 have lower mortality.

3.E. Early Life Risk Factors

Above we emphasized the role that acute stress during wartime had on the later life health and economic outcomes. Part of the stress experience by soldiers were serious wounds and injuries, the consequences of which could follow soldiers in many areas of their lives. For instance, Lee (2008) finds that wartime wounds and illnesses significantly diminished the veterans' geographic mobility after the war. On the health front, Wilson (2003) found that soldiers hospitalized for infectious disease during the war had significantly higher rates of chronic respiratory disease later in life.

Other scholars have used the Early Indicators data to link pre-War factors to later life outcomes. For example, Su (2009) finds that risk exposures in early life, including season of birth, country of origin, residential region, city size, and height at enlistment influence mortality risk many decades after the war. Hong (2007) looks at specific risk factor—exposure to malaria—to study the health of soldiers during the war. Those soldiers with high exposure to malaria were significantly shorter at enlistment due to malnutrition and were more susceptible to infections during the war.

One of the earliest uses of UA data was to examine the determinants of height, which is a proxy for the culumulative impact of infectious disease and nutrition during childhood. This line of research on US soldiers was sparked by Margo and Steckel (1983), who identified the puzzling "antebellum paradox" of declining height in a period of economic growth.[1] Wilson and Pope (2003) find significant effects of urbanization, socioeconomic status, occupation, and migration history on the adult height of soldiers in the UA collection. More recently, Zehetmayer (2011) has found that height is positively correlated with proximity to protein-rich nutrients during childhood and with geographic mobility. Height can still be fruitfully explored with new data on African-American troops and with the augment collection of urban recruits, where it can be combined with the wide array of ecological variables present in the HUE data.

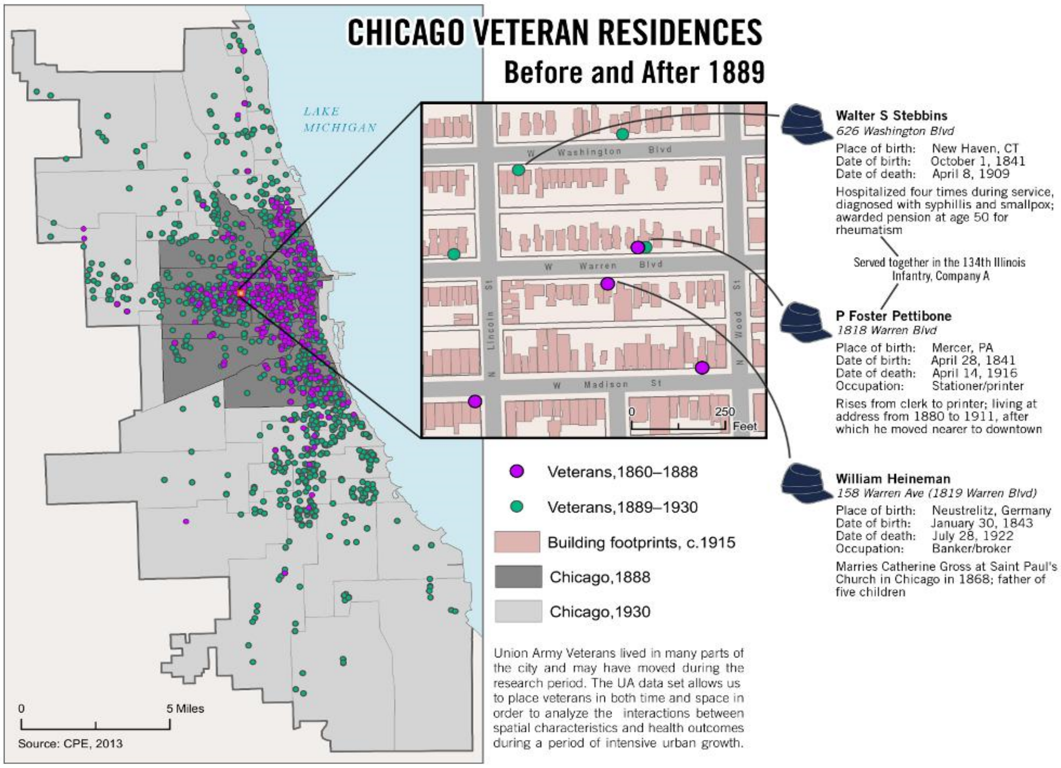
3.F. Social Networks

The *Early Indicators* collections have linked individuals in ways that are uncommon in modern datasets. Recruits were sampled as whole companies (usually 100-150 men), and companies were typically drawn heavily from a common location. Costa and Kahn (2010) have explored how these social networks affected the strength and cohesiveness of military units. Wilson, Burton and Villarreal (2014) are examining how these social networks can persist for many decades by following the residential clustering of company members in later life.

[1] See also Costa (1993), Costa and Steckel (1997) and Haines, Craig and Weiss (2000).

Because of the extensive geo-coding of the urban data, often to the street-level address, analyzing social networks can be facilitated by a geo-spatial analysis. Figure 3 shows a close-up of a Chicago neighborhood, for example, with two veterans living around the corner from one another who both served in the same Illinois company many years previous.

Figure 3: Veterans' Residential Patterns over Time, Chicago



CHICAGO VETERAN RESIDENCES
Before and After 1889

Walter S Stebbins
626 Washington Blvd
Place of birth: New Haven, CT
Date of birth: October 1, 1841
Date of death: April 8, 1909
Hospitalized four times during service, diagnosed with syphillis and smallpox; awarded pension at age 50 for rheumatism

Served together in the 134th Illinois Infantry, Company A

P Foster Pettibone
1818 Warren Blvd
Place of birth: Mercer, PA
Date of birth: April 28, 1841
Date of death: April 14, 1916
Occupation: Stationer/printer
Rises from clerk to printer; living at address from 1880 to 1911, after which he moved nearer to downtown

William Heineman
158 Warren Ave (1819 Warren Blvd)
Place of birth: Neustrelitz, Germany
Date of birth: January 30, 1843
Date of death: July 28, 1922
Occupation: Banker/broker
Marries Catherine Gross at Saint Paul's Church in Chicago in 1868; father of five children

Union Army Veterans lived in many parts of the city and may have moved during the research period. The UA data set allows us to place veterans in both time and space in order to analyze the interactions between spatial characteristics and health outcomes during a period of intensive urban growth.

Source: CPE, 2013

Family ties are also an important part of the veterans' network that can be exploited in the collection. Sizable pieces of the veterans' lives are obtained from linkage to the US federal censuses from 1850-1930, allowing information about both childhood family ties and living arrangements in later life. Socioeconomic information at the family level can be used to study the family-level effects on life outcomes. Though not yet part of the *Early Indicators* project, census records could be used to construct neighbor-hood level variables such as ethnic concentration, occupation distributions or wealth inequality.

3.G. Elderly Living Arrangements

Census collections allow researchers to study the lives of veterans in the context of the families and households in which they lived. By 1900, most veterans were getting pension support and they were, roughly, in the age interval 55-65—an ideal point to begin studying various aspects of the aging process related to living arrangements.

The table below gives one perspective on elderly living arrangements for Union Army veterans. The census enumerator indicated what member of each household was the "head," and each additional household member was identified by their relationship to the household head. As can be seen from Table 2, most veterans are listed as head, a percentage that declines gradually over the next two decades. Unfortunately, because the white sample was collected before the 1920 census was accessible, only white companies from Indiana and Wisconsin that were completed at a later date are included in this analysis.

Table 2: Living Arrangements among Veterans, by Race and Year

	1900		**1910**		**1920**	
Relationship to Head	Whites	Blacks	Whites	Blacks	Whites*	Blacks
Head	88.7%	88.8%	82.0%	83.3%	71.5%	77.9%
Parent	2.5%	1.8%	8.2%	3.6%	16.8%	8.7%
Spouse or Relative	2.0%	1.9%	2.0%	3.4%	2.8%	4.3%
Non-Relative	4.5%	6.7%	5.7%	7.6%	3.8%	5.7%
Alive but not linked to Census	2.3%	0.8%	2.2%	2.2%	5.0%	3.4%
N	12,737	4,201	7,567	2,600	499	960
Average Age (Census Age)	60.3	59.8	69.0	68.9	76.8	75.9
Age Range:	50-79		60-89		70-99	

* White sample in 1920 includes only veterans from Indiana and Wisconsin

The most striking feature of this very preliminary analysis, in our view, is the similarity between black and white living arrangements of elderly soldiers. In both 1900 and 1910, blacks and whites were household heads at very similar rates. The likelihood of living as a parent of the household head is, in 1910, 8.2% for whites and 3.6% for blacks, and that gap grows to almost 8 percentage points by 1920. It is likely the case that black veterans were more likely to have adult children who were less likely to live independently, possibly for economic reasons.

This analysis does not capture the rate at which white and black veterans may have lived with adult children as household members (in other words, what percentage of the veterans who were household heads had adult children or grandchildren living with them). Future work includes bringing more of this detail into the analysis, as well as information on the veterans' health, disability, and pension support. Of course bringing in national comparison data from the IPUMS data will allow comparisons between veteran households and the general population.

3.H. Chronic Disease Epidemiology

Research on the secular decline of chronic conditions and disabilities have been published with earlier versions of the data that have re-shaped how scholars think about the epidemiological transition (Fogel & Costa, 1997; Costa, 2000; Wilson, Burton, & Howell, 2005). In addition, the data collection remains a rich source of information for studying specific chronic diseases. Past work includes an examination of arteriosclerosis (Costa, Helmchen, & Wilson, 2007), respiratory disease (Wilson, 2003), and hernias (Song & Nguyen, 2003). Noymer (2009) uses the UA data to test whether tuberculosis was a risk factor for influenza epidemic, thereby reducing the prevalence of tuberculosis in the population.

These examples above uses only a small subset of the available data with which to study chronic disease epidemiology across the life course. Researchers on the *Early Indicators* team have worked for years to develop a system of standardizing the findings from the medical examinations that allow for detailed study of the epidemiology of dozens of chronic conditions in later life. Of course an enormous advantage of the UA collections is that risk factors for those conditions can be identified back to the early life experiences of the veterans.

There have been studies looking at the overall burden of chronic diseases and disabilities using the UA collection. Costa (2000, 2002) and Fogel and Costa (1997) document and explore the widespread decline in chronic disease rates and the associated decline in functional limitations. Costa (2000) emphasizes the role that the decline in manual labor and exposure to infectious disease as primary contributors to the decline in condition prevalence over a wide number of conditions. Wilson, Burton and Howell (2005) emphasize the importance of the decline in non-fatal conditions that in the 19th and early 20th century were highly prevalence and disabling but today seldom lead to disability; these include hernias, varicose veins, and a large number of gastrointestinal orders.

3.I. Intergenerational Forces

As noted above, family connections are an important aspect of the data collections. A recent renewal of funding by the NIA will involve opportunities for significant intergenerational analysis. Researchers have begun the process of linking the census and death information for the children of the Union Army veterans, both white and black. The new collection will also significantly increase the amount of data available on women, as the project collects that same data for both the sons and daughters of the veterans. New electronic search tools also facilitate higher find rates for women than have been available in the past.

We will thus be able to follow individuals across three generations and study the factors in the lives of the veterans which affect life outcomes into the next generation. For instance, from the 1850 and 1860 censuses, it is frequently possible to find the occupations of the soldiers' fathers, and the new data will give the occupations of their children all the way through the 1940 census. We know of no other data source that allows for the systematic study of multi-generational social and how that mobility intersects with health and longevity, not to mention other social factors such as race or immigration history.

As an illustration, Table 3 below combines data from the POW sample and a small VCC pilot done on white soldiers in the main sample. The table shows the occupation of veterans' sons as a function of age and birth order. In this table, first-born sons in their 20s are less likely to start their adult lives as laborers and more likely to be in professions or skilled trades. At older ages (30-39) first sons are even more likely to be in professions and less likely to be farmers. Sample sizes are as yet much too small to draw any conclusions, but one can envision using the ultimate VCC sample, once it has been collected, to study a host of demographic and economic factors related to social mobility in an intergenerational context.

Table 3: Occupation of Veterans' Sons, by Birth Order

	Age in 1900: 20-29			**Age in 1900: 30-39**		
Occupation in 1900	1st Son	2nd Son	3rd Son	1st Son	2nd Son	3rd Son
Laborer	33%	41%	43%	29%	25%	29%
Farmer (owner/operator)	21%	18%	23%	20%	29%	31%
Trade/Craft/Clerical	40%	36%	32%	40%	40%	36%
Manager/Professional	6%	5%	3%	11%	6%	4%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*
N=	349	253	160	159	164	107

4. Persistent Challenges and Continuing Progress

The Union Army collections are a vast and complex set of data covering a dynamic century of economic growth, industrialization, urbanization, and technological innovation. These features make the data both immensely valuable for research and immensely challenging to understand

and use appropriately. Any researcher using the collections needs to be aware of the significant challenges involved in avoiding the many sources of bias that can arise when using the data.

4.A. Understanding and Accounting for Bias

A thorough discussion of potential biases in the Union Army collections can be found on the *uadata.org* website and, to a large extent, in the many published papers using the data. In this section we highlight, in broad terms, some of the key issues that researchers need to be mindful of when undertaking analysis with the Union Army collections.

In an ideal situation, data on all sample members is generated by the same data generating process with parameters that depend only on factors that all sample individuals are at risk to experience at each point in time. The Union Army collections deviate from this ideal in numerous ways.   We highlight some of these in the following paragraphs.

1. Bias due to event-generated data

As noted above, the data collection captures information across the soldier's life course. But the existence of data at one point in time often depends on events that happen at a *future* date, which creates significant survival bias. The most ubiquitous source of such survival bias is caused by the event of applying for a Union Army pension. The pension application consists of a thick file of documents that describe life for many decades *before* the pension, including experiences during the war. Thus, some of the data in the collections during early life, during the war, and after the war, exists because the veteran lived to enter the pension, often much later in life.

The obvious complication in conducting analysis is that the data on two identical soldiers at a point in time, say 1870, will be very different if one of those soldiers survived to apply for a pension at a later date, say 1880, but the other did not. The pension also contains demographic and economic information, such as marriage and family data and occupation for the period covered between the war and the time of pension application. If the veteran dies without applying for this pension this information is unavailable, even though he may have lived for many decades beyond the end of the war.

One particularly problematic source of bias is when information from the pension files is used to populate data fields describing the wartime experience, such as claims of injuries or illnesses. In the main UAV file, data that comes from military files and data coming from later pension applications is not distinguished. However, special files have been created that give the information from CMSRs and CMRs without using any information from the pension application. Researchers needing an unbiased view of the wartime experience should utilize these important auxiliary files.

Surviving to participate in the pension increases the likelihood of linking the veteran to the 1850 and 1860 censuses because information obtained from the pension files and from the 1900 and 1910 censuses is very useful in obtaining the census records from the veteran's childhood. Thus an unbiased analysis of early life effects on later life outcomes is best conducted by conditioning on survival to a point in time where most veterans would have entered the pension. If one wanted to address, for instance, the effect of early life environment on, say, survival during the war, the researcher must account for the fact that the early life variables are much less likely to exist in the data for soldiers who died during the war and, therefore, never obtained a pension. Thus, the collection methods used to make linkages to the census impose serious limitations on how the early life census data can be used for research.

Another significant event that further complicates the nature of survival bias was the liberalization of the pension system in 1890. Prior to that point, veterans had to make an argument that their disability was "war-related." But under the 1890 law, any disability, regardless of cause, that limited the capacity to perform manual labor was eligible for pension support. This caused an immediate and significant increase in pension enrollment, which in turn generated a host of retrospective data of the kind we have been discussing. One important complication of this change for research, is any time series of disease or disability prevalence calculated from the medical data in the pension files must account for the difference in eligibility (and hence existence of the data) that occurred in 1890.

2. Variation and evolution in census linkage rates

Much of the important demographic information on the veterans' lives comes from the census records. An unavoidable problem in using this data is that linkage to the census is not random; it is highly correlated with factors that may, in turn, be correlated with the life outcome of interest. For example, urban environments were known to be much less healthy than rural ones, but soldiers in urban environments are also much harder to identify in the census because census linkage is difficult in urban areas. Place of birth is another variable that may be correlated with both the linkage rate and the outcome of interest.

The past decade has seen an explosion of new internet-based search tools available through sites such as *Ancestry.com*. These tools are now the primary method used to link veterans to both census records and vital records. This has led to much higher linkage rates for the later collections, namely the USCT, Urban, Andersonville and Oldest-Old samples. This increased linkage rate is very good news for users of those samples but a problem for researchers comparing those soldiers with the main Union Army collection, which was collected without those tools.

3. Historical change

The pension data cover decades of time in which advances in technology, medical knowledge and public health were occurring. These changes are reflected in the medical records found in the pension files. A typical physical examination record from 1870 looks vastly different than a typical examination record fifty years later in 1920. The increase in knowledge and professionalization of the examiners is readily apparent from even a casual observation. Such a transformation does not necessarily cause bias, but researchers must be careful in constructing time series of health outcomes from the pension files.

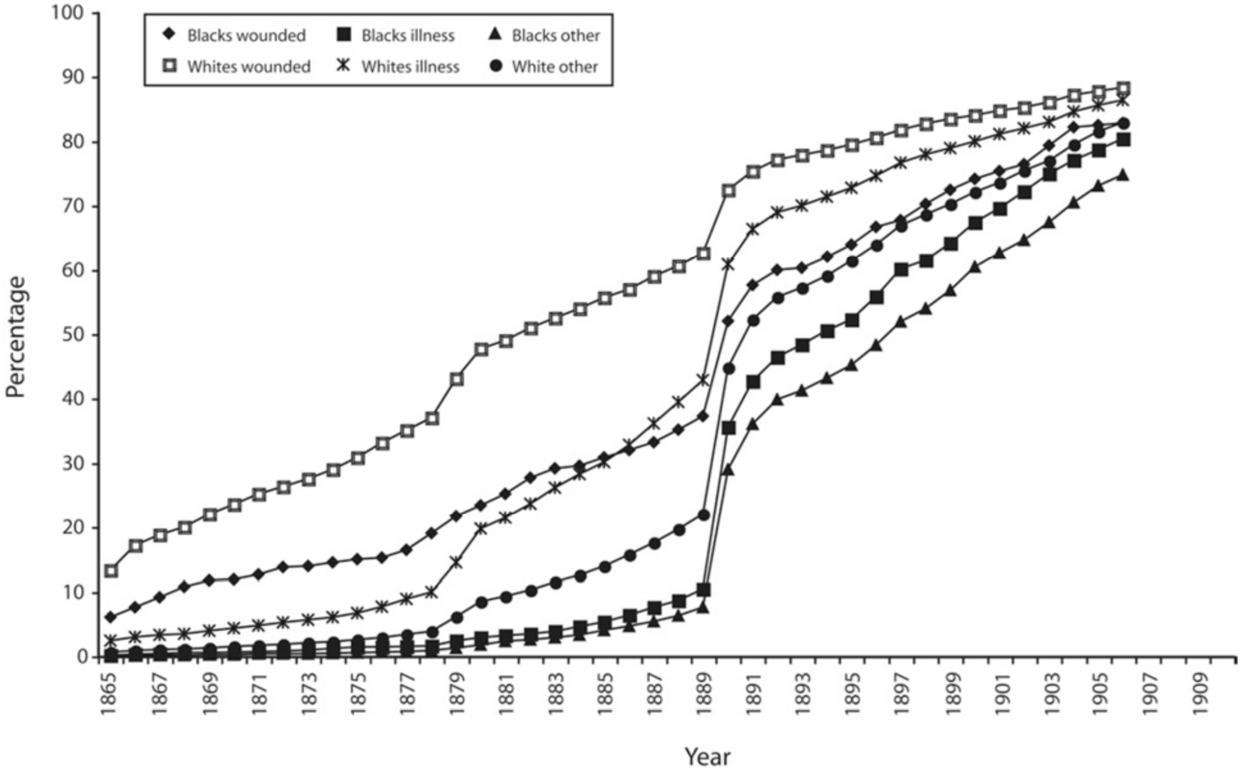
4. Administrative racial prejudice

One of the exciting new features of the Union Army collections is the large, new collection of records on African-American soldiers in the USCT. One of the first things many researchers will want to do is make comparisons between white and black soldiers. This, indeed, is one of the primary motivating goals behind the collection of the USCT data, but such comparisons are likely to hinge on underlying discrimination faced by black soldiers both during the war and after in the pension system.

Wilson (2010) provides a detailed discussion of prejudice in the pension system that was very effective at keeping black soldiers at a lower rate of participation. Part of this lower participation is due to the fact that blacks were less likely to be injured in battles than white soldiers, but a large part of the differential is due to discrimination. The figure below shows Wilson's estimated enrollment in the pension by race and by wartime medical history. Controlling for evidence in the military records, by the time the 1890 Act was passed whites had significantly higher enrollment than blacks in each category. Pension enrollment rates for those with a war wound were about 60% for whites, but only 35% for blacks. For those with an illness but no wound, whites were at 43% and blacks were at 10%. And for those without any illness or wound, whites were at 22% and blacks were at 8%.

In the post reconstruction period, it is easy to see from Figure 4 the informal liberalization of the pension system that was occurring between 1879 and 1889 prior to the formal liberalization in 1890. Starting with the passage of the 1879 Arrears Act, which allowed back payments of pension support, this was a period of intensive political activism on behalf of veterans. But blacks did not share equally from this liberalization. Their claims of war-related disabilities were rejected at higher rates, and they had a harder time gathering documentation that allowed them to verify their identities and their service. This was due to their status as slaves prior to service and to treatment during the war.

Figure 4: Pension Enrollment Proportions, By Race and Wartime Medical History



Source: Wilson (2010)

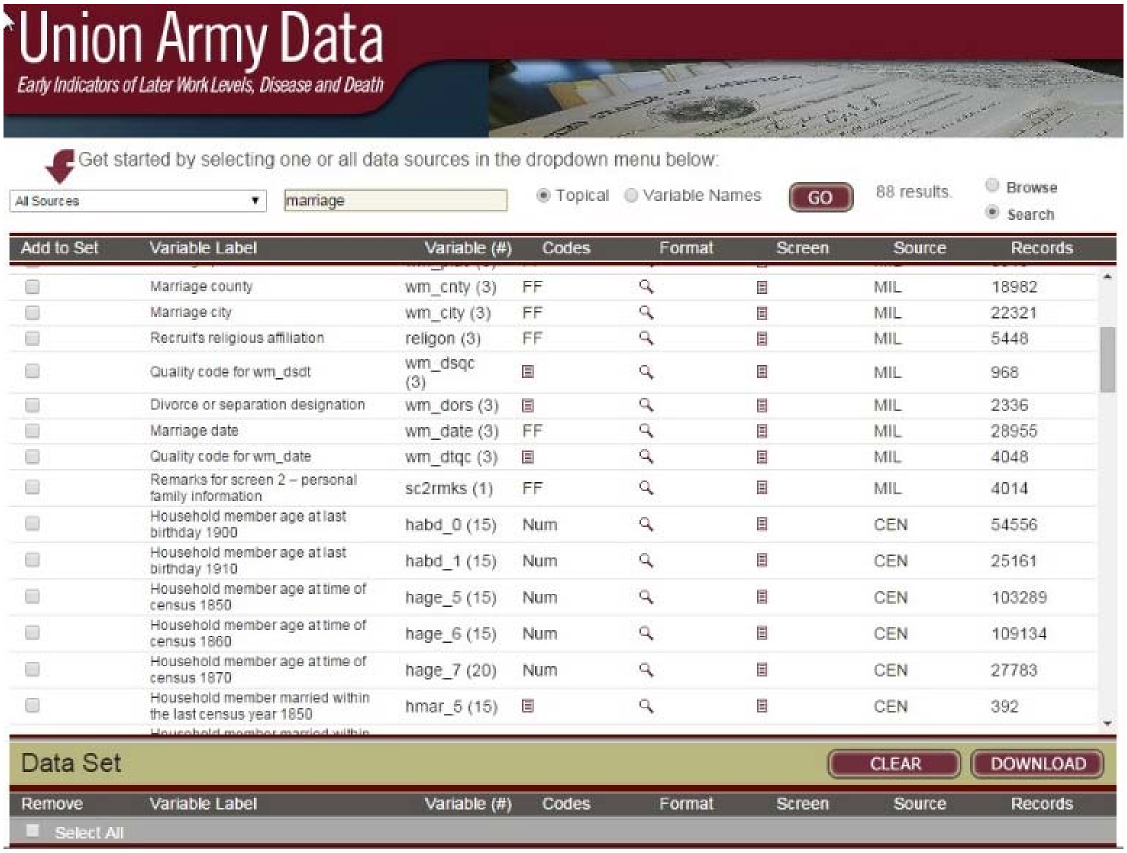
In the *Early Indicators* data, 18.7% of the black sample died from disease while in service, compared to only 9.6% of whites. Yet black troops were less likely to be sent to the hospital for illness than white troops. Scholars have argued that white medical officers “accused blacks of feigning sickness in much the same way that masters and overseers accused slaves of shirking work. They mistreated, abused, overworked, or neglected such soldiers, thereby contributing to further deterioration of their health” (Berlin, 1998, 636). Humphreys (2008) notes that black regiments were also understaffed compared to white regiments. The short-term impact of this discrimination was higher disease mortality during the war. The long-term impact was that blacks lacked documentation of their wartime illnesses, leading to a greater difficulty in claiming pension support in the period of informal liberalization.

In sum, the Union Army collections allow for detailed comparisons of the health of aging veterans, both black and white. But the blatant discrimination in the pension system is a fact that researchers must grapple with in making black-white comparisons using the pensions’ medical data.

4.B. User Services

Early Indicators researchers have put extensive work into easing the challenging task of using the Union Army collections. There is no short-cut to carefully studying the accompanying documentation, including sections on sample design and on the bias issues discussed above. But there are tools that can flatten the learning curve.

Figure 5: Union Army Data Searchable Extraction System



Get started by selecting one or all data sources in the dropdown menu below:

All Sources ▾ | marriage | ● Topical ○ Variable Names | GO | 88 results. | ○ Browse ● Search

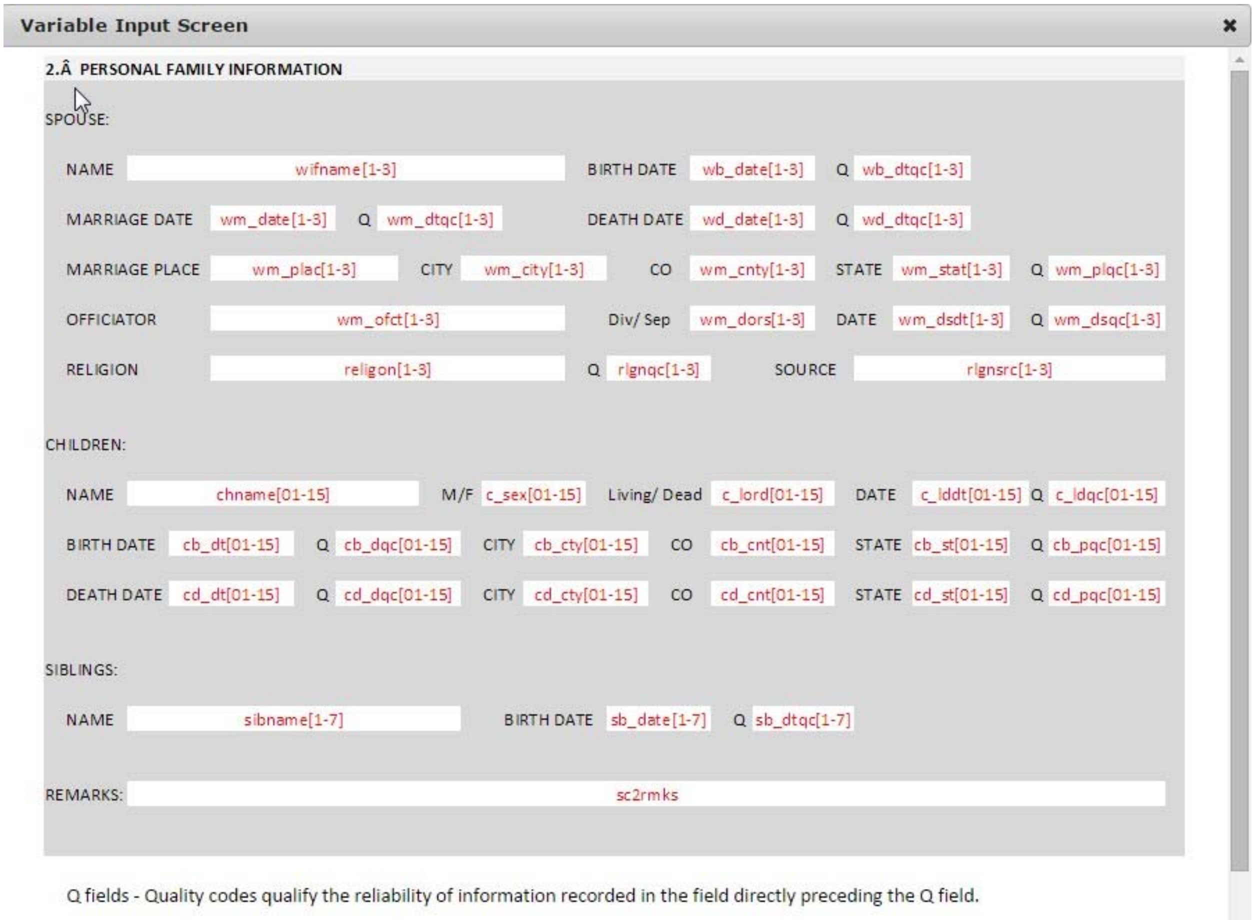
Add to Set	Variable Label	Variable (#)	Codes	Format	Screen	Source	Records
□	Marriage county	wm_cnty (3)	FF	🔍	▤	MIL	18982
□	Marriage city	wm_city (3)	FF	🔍	▤	MIL	22321
□	Recruit's religious affiliation	religon (3)	FF	🔍	▤	MIL	5448
□	Quality code for wm_dsdt	wm_dsqc (3)	▤	🔍	▤	MIL	968
□	Divorce or separation designation	wm_dors (3)	▤	🔍	▤	MIL	2336
□	Marriage date	wm_date (3)	FF	🔍	▤	MIL	28955
□	Quality code for wm_date	wm_dtqc (3)	▤	🔍	▤	MIL	4048
□	Remarks for screen 2 – personal family information	sc2rmks (1)	FF	🔍	▤	MIL	4014
□	Household member age at last birthday 1900	habd_0 (15)	Num	🔍	▤	CEN	54556
□	Household member age at last birthday 1910	habd_1 (15)	Num	🔍	▤	CEN	25161
□	Household member age at time of census 1850	hage_5 (15)	Num	🔍	▤	CEN	103289
□	Household member age at time of census 1860	hage_6 (15)	Num	🔍	▤	CEN	109134
□	Household member age at time of census 1870	hage_7 (20)	Num	🔍	▤	CEN	27783
□	Household member married within the last census year 1850	hmar_5 (15)	▤	🔍	▤	CEN	392

Data Set | CLEAR | DOWNLOAD

Remove	Variable Label	Variable (#)	Codes	Format	Screen	Source	Records
□ Select All							

First, the data collections and documentation can be downloaded in bulk for users accustomed to tackling large datasets. But the new extraction system can be a valuable tool for the new researcher. The extraction system is organized topically, allowing a user to hone in on a set of variables from all data sources related to the topic of interest. For example, Figure5X shows a subset of the results from a search on the topic of marriage, illustrating both census and pension-based variablesThe columns shown in the extraction system provide the user with quick links to the codes and formatting used for each variable. Each variable is also linked to the data collection screen from which it comes. This is an invaluable tool for learning how variables relate to each other and the user can see, in this example, the marriage information available on Screen 2.A., as well as a variety of other family information, including the variable names.

Figure 6: Example of Data Inputting Screen



Variable Input Screen

2.Â PERSONAL FAMILY INFORMATION

SPOUSE:

NAME wifname[1-3] BIRTH DATE wb_date[1-3] Q wb_dtqc[1-3]

MARRIAGE DATE wm_date[1-3] Q wm_dtqc[1-3] DEATH DATE wd_date[1-3] Q wd_dtqc[1-3]

MARRIAGE PLACE wm_plac[1-3] CITY wm_city[1-3] CO wm_cnty[1-3] STATE wm_stat[1-3] Q wm_plqc[1-3]

OFFICIATOR wm_ofct[1-3] Div/ Sep wm_dors[1-3] DATE wm_dsdt[1-3] Q wm_dsqc[1-3]

RELIGION religon[1-3] Q rlgnqc[1-3] SOURCE rlgnsrc[1-3]

CHILDREN:

NAME chname[01-15] M/F c_sex[01-15] Living/ Dead c_lord[01-15] DATE c_lddt[01-15] Q c_ldqc[01-15]

BIRTH DATE cb_dt[01-15] Q cb_dqc[01-15] CITY cb_cty[01-15] CO cb_cnt[01-15] STATE cb_st[01-15] Q cb_pqc[01-15]

DEATH DATE cd_dt[01-15] Q cd_dqc[01-15] CITY cd_cty[01-15] CO cd_cnt[01-15] STATE cd_st[01-15] Q cd_pqc[01-15]

SIBLINGS:

NAME sibname[1-7] BIRTH DATE sb_date[1-7] Q sb_dtqc[1-7]

REMARKS: sc2rmks

Q fields - Quality codes qualify the reliability of information recorded in the field directly preceding the Q field.

The data extraction system is a helpful tool for users to quickly obtain needed variables. After selecting the variables they want, users can indicate on the export page which collection they want to draw from (Whites, Blacks, Urban, etc.), and how they want the data formatted (Excel, SAS, Stata, etc.). Each export comes with a customized set of online documentation that provides general background information as well as documentation that is unique to the exported variables.

But the primary function of the extraction system is not extraction, but teaching. The topical based approach to the data, which can be browsed or searched, is designed to help users understand the collections and to ease the path of entry into these complex data collections. In addition, the website provides a large set of topical user guides that provide the user with extensive detail on the variables and the sample design and collection methods used to obtain them. Personal assistance from the Early Indicators team is also readily available, and the staff will respond to queries as soon as possible.

5. Conclusion

The Union Army data collections began accumulating nearly 30 years ago. And since that time these soldiers have been contributing to our understanding of health, mortality, economic mobility, and a variety of scientific questions across many disciplines. By this measure, the UA collections are old news.

But we have tried to illustrate with this essay that many vital questions related to the human experience that might be illuminated by the UA collections are still to be explored and answered. Many of these questions are best served by having data on a large set of individuals over their entire lifetimes. Some of the key forces that seem to shape human health and flourishing—poverty, stress, inequality, discrimination, family dissolution, and environmental risks—can be addressed with this remarkable and growing set of life histories.

References

Berlin, I., J.P. Reidy, and I.S. Rowland. 1998. *Freedom's soldiers: The Black Military Experience in the Civil War*. Cambridge, England: Cambridge University Press.

Canavese, Paula and Robert Fogel. 2010. Arthritis: Changes in its Prevalence during the Nineteenth and Twentieth Centuries." in David M. Cutler and David A. Wise (eds.) *The Causes and Consequences of Declining Disability among the Elderly.* Chicago: NBER and University of Chicago Press.

Costa, Dora L. 1993. Height, wealth, and disease among the native born in the rural antebellum North. *Social Science History* 17: 355-83.

Costa, Dora L. 2000. "Understanding the 20th Century Decline in Chronic Conditions Among Older Men." *Demography* 37:53-72.

Costa, Dora L. 2002. Changing Chronic Disease Rates and Long-term Declines in Functional Limitations Among Older Men. Demography 39:119-138.

Costa, Dora L. 2010. “Pensions and Retirement Among Black Union Army Veterans.” *Journal of Economic History*. 70(3): 567-92.

Costa, Dora L. 2012. “Scarring and Mortality Selection Among Civil War POWs: A Long-Term Mortality, Morbidity, and Socioeconomic Follow-Up” *Demography.* November 2012. 49(4):1185-206.

Costa, Dora L., Lorens Helmchen and Sven E. Wilson. 2007. "Race, Infectious Disease, and Arteriosclerosis." *Proceedings of the National Academy of Sciences* 104: 13219-13224.

Costa, Dora L. and Matthew Kahn. 2010. “Health, Wartime Stress, and Unit Cohesion: Evidence from Union Army Veterans.” *Demography* 47(1): 45-66.

Costa DL, Kahn ME. 2007. Surviving Andersonville: The benefits of social networks in POW camps. *American Economic Review.* 97(4): 1467-1487

Costa, Dora L., and Richard Steckel. 1997. Long-term trends in health, welfare, and economic growth in the United States. In *Health and welfare during industrialization,* ed. Richard H. Steckel and Roderick Floud, 47-90. Chicago: University of Chicago Press.

Fogel, Robert W., and Dora L. Costa. 1997. “A Theory of Technophysio Evolution, With Some Implications for Forecasting Population, Health Care Costs, and Pension Costs.” Demography 34: 49-66.

Haines, Michael R., Lee A. Craig, and Thomas Weiss. 2000. “Development, health, nutrition, and mortality: The case of the "antebellum puzzle" in the United States.” NBER Working Paper no. H0130. Cambridge, Mass.: National Bureau of Economic Research

Hong SC. 2007. The burden of early exposure to malaria in the United States, 1850-1860: Malnutrition and immune disorders. *Journal of Economic History* 67(4): 1001-1035.

Humphreys, Margaret. 2008. *Intensely Human: The Health of the Black Soldier in the American Civil War*. Baltimore: The Johns Hopkins University Press.

Lee C. 2008. Health, information, and migration: Geographic mobility of Union Army veterans, 1860-1880. *Journal of Economic History* 68(3): 862-899.

Noymer A. 2009. Testing the influenza-tuberculosis selective mortality hypothesis with Union Army data. *Social Science Medicine* 68(9): 1599-1608

Margo, Robert A., and Richard H. Steckel. 1983. Heights of native-born whites during the antebellum period. *Journal of Economic History* 43 (1): 167-74

Pizarro, Judith, Roxane Cohen Silver and JoAnn Prause. 2006. "Physical and mental health costs of traumatic war experiences among civil war veterans." *Archives of General Psychiatry, 63*, 193-200.

Salisbury, Laura. 2014. "Women's Income and Marriage Markets in the United States: Evidence from the Civil War Pension." NBER Working Paper No. 20201.

Song, Chen and Louis L. Nguyen. 2003. "The Effect of Hernias on Labor Force Participation of Union Army Veterans." in Dora Costa (ed.) *Health and Labor Force Participation over the Life Cycle: Evidence from the Pas.,* 2003. Chicago: NBER and University of Chicago Press.

Su D. 2009. Risk exposure in early life and mortality at older ages: Evidence from Union Army veterans. *Population and Development Review.* 35(2): 275-295.

Wilson, Sven E. 2003. "The Prevalence of Chronic Respiratory Disease in the Industrial Era: The United States, 1895-1910," in Dora Costa (ed.) *Health and Labor Force Participation over the Life Cycle: Evidence from the Past.* Chicago: NBER and University of Chicago Press.

Wilson, Sven E. 2010. "Prejudice and Policy: Racial Discrimination in the Union Army Disability Pension System, 1865-1906." *American Journal of Public Health* 100 : S56-S65.

Wilson, Sven E., Joseph Burton and Benjamin Howell. 2005. "Work and the Disability Transition in Twentieth Century America." *NBER Working Paper* No. 11036.

Wilson, Sven E. and Clayne Pope. 2003. "The Height of Union Army Recruits: Family and Community Influences" in Dora Costa (ed.) *Health and Labor Force Participation over the Life Cycle: Evidence from the Past.* Chicago: NBER and University of Chicago Press.

Wilson, Sven E., Carlos Villarreal and Joseph Burton. 2014. "The Long-term Persistence of Social Capital. Residential Clustering among Union Army Veterans." Unpublished manuscript.

Zehetmayer M. 2011. The continuation of the antebellum puzzle: Stature in the US, 1847-1894. *European Review of Economic History* 15(2): 313-327.