Understanding the Role of Adolescent Housing Residence on Adverse Childhood Experiences and Outcomes of Chronic Disease Risk

Data Linkage Report

U.S. Department of Housing and Urban Development | Office of Policy Development and Research

DISCLAIMER

The contents of this report are the views of the contractor and do not necessarily reflect the views or policies of the U.S. Department of Housing and Urban Development or the U.S. Government.

Understanding the Role of Adolescent Housing Residence on Adverse Childhood Experiences and Outcomes of Chronic Disease Risk

Data Linkage Report

Prepared by:
Atticus Jaramillo

Dr. Jon Hussey
Dr. Michael D. Webb
Tim Monbureau
Maria Marrufo

Add Health

November 2020

Table of Contents

Table of Exhibits

Executive Summary

To better understand how residence in the U.S. Department of Housing and Urban Development (HUD)-assisted housing affects future outcomes, HUD is supporting the linkage of HUD administrative data to existing national cohort studies. This report documents the linkage of HUD administrative records to data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), the largest and most comprehensive longitudinal study of adolescent health ever conducted in the United States. The report also outlines summary statistics on linked sample members' length of residence in HUD-assisted housing, the number of spells in HUD-assisted housing (termed "episodes"), and concurrency between HUD-assisted housing residence and Add Health survey participation. The linkage of those two data sources will provide the research community with an unprecedented opportunity to examine a wide range of social, behavioral, economic, and health outcomes—from adolescence into middle adulthood—associated with assisted-housing residence.

Methods

More than 70.6 million HUD individual resident (member) and household-level records were searched to identify Add Health study participants. All members of the original Add Health longitudinal cohort (n=20,745) with non-missing data on the partial identifiers were linkage eligible. Add Health and HUD records were linked probabilistically using eight partial identifiers. To ensure confidentiality, the authors suppressed summary statistics for outcomes that pertain only to 10 or fewer linked sample members.

Highlights

- The data linkage produced a single hierarchical (long-format) data file on 1,159 unique Add Health respondents who lived in HUD-assisted housing at some point between 1995 and 2017. Throughout this report, those 1,159 Add Health respondents are referred to as "linked sample members."
- Of linked sample members, 51 percent had two or fewer episodes of residence in HUD-assisted housing, and 69 percent had three or fewer episodes.[1]
- A plurality of linked episodes (43.3 percent) were for the Housing Choice Voucher (HCV) program, followed by the public housing (27.5 percent), Project-Based Section 8 (26.3 percent), and other multifamily (2.9 percent) programs, respectively.
- The mean length of linked sample members' first housing episode was roughly 1.5 years. Subsequent episodes' means ranged from 1.5 to 2 years. Across all episodes, however, the standard deviation for episode length is roughly equivalent to the mean, indicating significant variability in episode length.
- On average, 30–40 percent of linked sample members who resided in HUD-assisted housing when Add Health was in the field responded to a given Add Health survey.

[1] The term "episodes" refers to the specific time periods when linked sample members lived in HUD-assisted housing. Episodes were calculated based on HUD transaction records.

Introduction

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is an ongoing cohort study of a nationally representative sample of 20,745 U.S. adolescents in grades 7 through 12 in the 1994–1995 school year who have been followed for 25 years, with surveys in 1994–1995 (Wave I), 1996 (Wave II), 2001–2002 (Wave III), 2008–2009 (Wave IV), and 2016–2018 (Wave V).[2] Originally designed in response to a congressional mandate for a national study of adolescent health, Add Health has evolved into a unique national data resource for studying life course influences on health and well-being from adolescence into middle adulthood. Since its inception in 1994, Add Health has been funded through five program project grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cofunding from more than 20 other federal agencies and foundations. Add Health is the largest, most comprehensive longitudinal study of adolescents ever conducted in the United States.

In addition to the core surveys administered at each wave of data collection, the scientific value of the Add Health study has been significantly increased through the addition of many other sources of data on the cohort participants. As part of the Wave IV and V program projects, for example, anthropometric (height; weight), cardiovascular (blood pressure), metabolic (HbA1c), and inflammatory biomarkers were collected. Multiple investigators have also obtained independent funding to link a wealth of supplemental data to the Add Health cohort database, including extensive measures of neighborhood- and county-level characteristics and high school transcripts.

This report describes our effort to supplement the Add Health database by identifying and linking administrative records from the U.S. Department of Housing and Urban Development (HUD) for any Add Health cohort member who resided in HUD-assisted housing between 1995 and 2017.

These datasets were linked to facilitate a HUD-funded research project, housed at the University of North Carolina-Chapel Hill, investigating relationships between residence in HUD-assisted housing as a teen, exposure to adverse childhood experiences (ACEs), and chronic disease later in life. In addition to informing this specific research project, the linked data set provides the research community with an unprecedented opportunity to study how residence in HUD-assisted housing affects health and well-being into mid-adulthood.

The remainder of the report proceeds as follows. The next section summarizes the two data sources used in the linkage and provides some additional information on HUD programs. The following section details the procedures used to approve the linkage and determine eligible members of the linkage and explains the linkage process. The report then describes the linked dataset in addition to its storage and restrictions placed on its access.

[2] Wave V data were released following the data linkage.

Data Sources

This section of the report discusses the data sources used in the linkage process. The report first describes the National Longitudinal Study of Adolescent to Adult Health (Add Health) and then provides a high-level overview of HUD programs before discussing HUD data collection procedures and relevant data sets.

Add Health

Add Health is an ongoing prospective cohort study that has followed a nationally representative sample of U.S. adolescents into adulthood. The study began in 1994–1995 with adolescents in grades 7 through 12 and has followed these respondents through adolescence and the transition to adulthood, with five completed waves of data collection as of 2020. The original goal of the survey was to explain the causes of adolescent health and health behavior, with emphasis on the multiple contexts of adolescent life, including school, peer groups, neighborhood, and family. As the cohort has aged, the study objectives have been expanded to include a focus on how adolescent and early adulthood experiences, behaviors, and social contexts influence well-being in mid-adulthood, using an integrative approach that combines the social, behavioral, and biomedical sciences. In addition to anthropometric, blood-based, genetic, and self-reported health data, the Add Health database—available in public-use and restricted-use versions—includes social, economic, behavioral, demographic, and contextual data.

Add Health is based on a multistage, stratified, school-based, cluster-sampling design. A stratified sample of 80 high schools were selected, with probability proportional to size, from a comprehensive list of U.S. high schools. Schools were first stratified by size, type (public, parochial, private), region (Northeast, Midwest, South, West), urbanization (urban, suburban, rural), and racial/ethnic composition (Thalji et al., 1997). For each high school, one feeder school (typically a middle school) was selected with probability proportional to its student contribution to the high school. Because some schools spanned grades 7 through 12, the sample contains 132 schools in 80 communities. School sizes varied from fewer than 100 students to more than 3,000 students. All 7th- through 12th-grade students in those 132 schools were eligible for the in-school questionnaire, which was completed by 90,118 students between September 1994 and April 1995.

Wave I of Add Health also included an in-home interview. The sampling frame for the in-home interview included students listed on their school enrollment roster and students not on the roster but who completed the in-school questionnaire. From that list, a probability sample of students were selected, including oversamples of Cuban, Puerto Rican, and Chinese adolescents; African-American adolescents with a college-graduate parent; and adolescents with disabilities. In addition, the sample included a genetic-relatedness oversample of twins, siblings, half-siblings, and unrelated adolescents living in the same household. Wave I in-home interviews were completed between April and December 1995 with 79 percent of sampled respondents, for a total sample size of 20,745 adolescents aged 12 to 19 years. An in-home interview also was completed with a parent, usually the resident mother.

The Wave I in-home sample of adolescents is the basis for all longitudinal followup interviewers. Three followup waves have been completed to date. Wave II was fielded in 1996, approximately 1 year after the baseline. Interviews were completed with 88 percent of sampled respondents, consisting primarily of Wave I respondents in grades 7 through 11, for a total Wave II sample size of 14,738. In 2001–2002, Wave III interviews were completed with 15,170 respondents, who at the time were 18 to 26 years of age, from the original probability sample, with a response rate of 77 percent. At Wave IV (2007–2008), interviews were completed with 15,701 respondents, ranging in age from 24 to 32 years (80-percent response rate). At Wave V (2016–2018), data collection was completed with 12,300 respondents, ages 31 to 42 (63-percent response rate). Survey data from this most recent wave were released to contractual users in December 2019.

Written parent or guardian consent and adolescent assent were obtained before the Wave I and II interviews. At Waves III, IV, and V, written consent was obtained from all respondents (then adults). A complete description of the Add Health study design and sample is available elsewhere (Harris et al., 2019).[3] All Add Health procedures and the present study were approved by the Public Health Institutional Review Board (IRB) at the University of North Carolina-Chapel Hill.

As with all school-based samples, potential bias due to the absence of dropouts is a concern with Add Health. Unlike most other school-based studies, however, the magnitude of this bias has been estimated in the Add Health sample. Udry and Chantala (2003) found that the bias in estimates of key risk behaviors due to the omission of those not enrolled in school is negligible when the population of interest is 15- to 19-year-old males or females. It is also important to note that the Add Health longitudinal cohort includes some high school dropouts (specifically, in-school respondents selected for the in-home survey who subsequently dropped out of school).

Potential bias due to nonresponse has been investigated. Kalsbeek, Morris, and Vaughn (2001) reported that bias due to nonresponse in Wave I and Wave II prevalence estimates of cigarette and alcohol use, depression, violent behavior, and other major health risks rarely exceeds 1 percent. At Wave III, Kalsbeek and colleagues expanded their analysis to evaluate nonresponse bias for 67 measures—including indicators of childhood socioeconomic status, health status, and health risk behaviors—and again found it to be minimal (less than 1 percent) in nearly all cases (Chantala, Kalsbeek, and Andraca, 2004).

HUD Programs

This section provides background information on affordable housing programs offered by HUD. It first provides a concise overview of the three primary programs through which HUD offers housing assistance. The section then reviews HUD administrative data, including the forms used to collect data and the systems on which the data are stored.

[3]Add Health website: https://addhealth.cpc.unc.edu/.

HUD offers housing assistance to low-income, very low-income, and extremely low-income families through various programs.[4] The largest of those program categories include Housing Choice Vouchers (HCVs),[5] public housing (PH), and multifamily (MF) programs. The former two are administered under HUD's Office of Public and Indian Housing, whereas the latter is administered under HUD's Office of Housing.

Another key distinction in HUD's housing assistance programs is whether assistance is project or tenant based. Project-based assistance entails tying the HUD subsidy to a particular housing unit or development and is most closely associated with the public housing and multifamily programs (although, as discussed later, a portion of the HCV program is also project based). With tenant-based assistance, the HUD subsidy is tied to a particular household; this assistance is most closely associated with the HCV program, with the exception of project-based HCV units.

In 2019, the most recent period for which data are available, HUD provided more than 5 million (5,035,000) units of subsidized housing nationwide that collectively served 9,439,919 individuals (HUD, n.d.d). The programmatic breakdown for those units is as follows:

- HCVs: 2,556,000 vouchers serving 5,249,000 individuals (As discussed in the following section, some of those vouchers are tenant based, whereas others are project based.)
- PH: 987,000 units serving 1,909,000 individuals
- MF: 1,491,000 units serving 2,282,000 individuals
    - Of those units, 1,290,000 are Project-Based Section 8; those units serve 2,063,000 individuals.
    - The remainder are in smaller MF programs, such as Section 202 and Section 811 (discussed in a following section; HUD, n.d.d).

Public Housing and the Housing Choice Voucher Program

Public housing authorities (PHAs) administer the HCV and PH programs locally, and they are responsible for screening tenants for both programs and maintaining waitlists for assistance. As of 2019, there are approximately 3,900 PHAs nationwide (NCHS OAE, 2019).[6]

---

[4] For housing assistance programs, HUD defines income limits as a percentage of area median income (AMI) for a family of four, with adjustments for household size. Low income means income not greater than 80 percent of AMI, very low income is not greater than 50 percent of AMI, and extremely low income is not more than 30 percent of AMI. Most HUD-assisted households are found in the latter two categories.

[5] HCV is often referred to (misleadingly) as "Section 8" because it was initially implemented under Section 8 of the Housing Act of 1937, although it was amended by and made much more widespread following the passage of the Housing and Community Development Act of 1974. Yet another HUD program is known as "Project-Based Section 8," which does not rely on vouchers.

[6] National Center for Health Statistics, Office of Analysis and Epidemiology. A Primer on HUD Programs and Associated Administrative Data. June 2019. Hyattsville, Maryland. https://www.cdc.gov/nchs/data-linkage/hud-methods.htm. See https://www.hud.gov/program_offices/public_indian_housing/pha/contacts for a list of all PHAs.

For both the PH and HCV programs, households' rents are set at roughly 30 percent of their income after adjusting for items such as childcare and medical expenses and utility allowances. Thus, households' rents may change during the year due to changes in either their income or their deductions. This income-based method of setting tenant rents is known as a "deep subsidy."

Although PHAs own and manage public housing buildings, since the mid-1990s, the number of units in the PH program has declined by 250,000 (CBPP, 2017), largely due to three factors:

- The HOPE VI (1993–2010) (HUD, n.d.a) and Choice Neighborhoods (2010–present) (Pendall and Hendey, 2013) programs, through which PHAs received funding to redevelop distressed public housing complexes, often into mixed-income communities. Those redevelopments often led to a net reduction in the number of PH units on the site, with some residents shifting their participation in HUD-assisted housing to the HCV program (Gress, Cho, and Joseph, 2017).
- The Rental Assistance Demonstration (RAD), through which PHAs may convert PH units to either project-based vouchers (PBVs) or project-based rental assistance (Hanlon, 2017).
- The general obsolescence of PH developments due to age and deferred maintenance, which has resulted in their demolition or disposition through HUD's Section 18 program or through the conversion of their assistance to HCVs through the Section 22 Voluntary Conversion program (HUD, n.d.e).

In contrast to PHAs' dual ownership and management roles for PH programs, PHAs simply administer HCVs; PHAs receive both voucher and administrative funding from HUD and disburse housing assistance payments to private-market landlords participating in the program. Within the HCV program, there are two types of vouchers offered: tenant-based vouchers (TBVs) and PBVs. Participants who receive TBVs have the ability to relocate with the voucher between leases, provided their new landlord accepts the voucher. Furthermore, TBV recipients may relocate to a different jurisdiction with their voucher, a process called "porting."[7]

PHAs may choose to convert some of their TBV allocations to the PBV program. In this case, the PHA enters into a contract with a building owner to reserve certain units for families qualifying for voucher assistance. Unlike TBVs, the voucher is tied to the unit (with TBVs, the voucher is tied to the participant). Families on the HCV waitlist are eligible to move into PBV units, and as with TBVs, the property owner has the final approval for whether a participant may move into a unit. After 1 year of

[7] 24 CFR 982.355 (a) When a family moves under portability (in accordance with 24 CFR 982.353(b)) to an area outside the initial PHA jurisdiction, another PHA (the "receiving PHA") must administer assistance for the family if a PHA with a tenant-based program has jurisdiction in the area where the unit is located.

living in a PBV unit, a participant may choose to relocate and will receive the next available TBV when one becomes available.

Multifamily Programs

Unlike PHA-administered programs such as PH, HCV, PBV, and Moderate Rehabilitation ("Mod Rehab"),[8] HUD directly administers multifamily programs and provides direct payments to participating property owners. For multifamily housing, HUD enters into contracts with property owners who, in turn, reserve at least four of their units for low-income families. These affordability restrictions typically limit the rent that landlords may charge for a unit and may include restrictions on the type of households who may occupy a unit—for instance, multifamily properties through the Section 811 program are restricted to adults with disabilities. The largest multifamily program assisted with deep subsidies is Project-Based Section 8, which serves 1.3 million of the total 1.5 million multifamily units.

Within the broad category of "multifamily" programs exist numerous smaller subprograms, some of which have been phased out or substantially amended over time. What multifamily programs have in common is that HUD has provided some form of shallow subsidy through Federal Housing Administration mortgage insurance, capital advances, or other lower cost financing in return for commitments to provide long-term rent affordability at low- and moderate-income levels. A number of multifamily programs, including most that are reported in HUD's tenant data system, include some type of rental assistance contract between HUD and the owner that provides deep subsidies so that rents are affordable even for extremely low-income households.[9]

Moving to Work

Moving to Work (MTW) is a HUD demonstration program that provides participating PHAs with flexibility to innovate in how they deliver housing assistance.[10] Enacted as part of the 1996 federal budget, MTW allows participating PHAs to waive or modify—with HUD approval—regulations in the 1937 Housing Act (such as income recertification policies, limits on project-basing units, and procedures for inspecting HCV clients' units). MTW also allows participating PHAs to combine several federal funding streams—that is, PH operating and capital funds, HCV operating and administrative funds—into

[8] Mod Rehab is a small and shrinking multifamily program administered by PHAs. For sizes of programs, see HUD (n.d.d). For greater detail on various multifamily programs, see NCHS (2019).

[9] Multifamily program types represented in the HUD tenant data systems discussed in this report include Project-Based Section 8; Section 811 Project Rental Assistance (PRA) Demonstration; Section 811 Project Rental Assistance Contract (PRAC); and Section 202 PRAC or PAC, as well as several programs under the heading of Other multifamily that were financed through mortgage financing vehicles no longer offered. Other multifamily programs include two shallow subsidy programs, Section 236 and Section 221(d)(3) Below Market Interest Rates (BMIR), as well as two deep subsidy programs, Rent Supplement and Rental Assistance Payments (RAP). HUD is converting Rent Supplement and RAP projects to Project-Based Section 8 through the Rental Assistance Demonstration.

[10] For greater detail on the MTW programs, including a timeline of participating PHAs and a list of MTW activities implemented across participating agencies, see Webb, Frescoln, and Rohe (2016).

a single, flexible account. One should note that MTW affects only PHAs and their PH and HCV programs, not all MF programs.

As of June 2020, 39 PHAs participate in the MTW demonstration, although Congress has authorized an additional 100 PHAs to join the demonstration within the next few years (HUD, n.d.c). Among the 39 current participants, some are very large PHAs (Baltimore, Chicago, Philadelphia, Seattle, and Washington, D.C.), whereas others are mid-sized agencies (Cambridge, MA; Charlotte; and Orlando) and still others are small agencies (Boulder, CO; Champaign County, IL; and Portage County, OH).

Specific programs or policies that PHAs implement though MTW are called "activities," and some of them affect data collection and reporting. For one, MTW agencies report data on a different HUD form, called HUD-50058 MTW (discussed in greater detail later in this document). Another of the most commonly implemented activities is to reduce the frequency of recertifications for some or all clients a PHA serves. Many MTW agencies now require senior clients and clients with disabilities to be recertified only every 2 or 3 years, and several require non-senior/disabled clients to be recertified only biennially. As a result, MTW PHAs may transmit tenant data to HUD on a less frequent basis, a factor that affected how episodes were created in the linked data set (discussed in a later section).

HUD Administrative Data

Except for MTW agencies, HUD-assisted housing programs primarily collect data for participants at two times: when a tenant initially moves into a unit and, at a minimum, annually thereafter. For the PH and HCV programs, PHAs can require participants to report any changes in income (either positive or negative) between annual recertifications (known as "interim recertifications"), as those income changes may trigger a recalculation in rent. Failure to report those changes may result in program termination. For MF programs, data collection occurs when a family moves into the unit and annually thereafter with interim recertifications similar to those for public housing and HCV programs (e.g., changes in family composition or income). MTW PHAs are an exception to this rule, however, and as discussed previously, may collect data less frequently than annually.

The data collected by PHAs and multifamily housing providers via administrative forms (mentioned below) include the following:

- Date of data collection
- Reason for data collection—e.g., program entry, income recertification, annual recertification, or end of program participation
- Geographic location of the housing unit
- Personal characteristics of everyone living in the housing units:
    - First and last name
    - Race/ethnicity
    - Sex
    - Citizenship status
    - Date of birth

    - Social Security number
    - Relationship to head of household
- Information on income and assets, including income sources, in addition to an estimate for the household's income in the next 12 months.

These data are collected and submitted to HUD on the following forms:

- HUD-50058 (the "Family Report): used by PHAs to collect data on households who participate in the HCV and PH programs. http://portal.hud.gov/hudportal/documents/huddoc?id=HUD50058.pdf .
- HUD-50058 MTW (the "MTW Family Report"): used by MTW PHAs to collect data on households who participate in any MTW program offered by a MTW PHA.[11] https://www.hud.gov/sites/documents/DOC_10236.pdf.
- HUD-50059 ("Owners Certification of Compliance with HUD's Tenant Eligibility and Rent Procedures"): used by MF housing providers to collect data on households participating in MF programs. https://www.hud.gov/sites/documents/50059.PDF.

Inventory Management System of the Public and Indian Housing (PIC) Information Center (IMS/PIC)
Since the mid-1990s, HUD has maintained a centralized repository of data on households participating in the PH and HCV programs. Before the early 2000s, PHAs submitted data for HCV and PH households to the Multifamily Tenant Characteristics System (MTCS). In the early 2000s, MTCS data were migrated to the Public and Indian Housing (PIH) Information Center (PIC); greater detail on that transition is provided later in this document. Some of the data used for the linkage described in this report—which spans years 1995 to 2017—come from MTCS, as HUD did not begin to transition to PIC until December 1999, and the transition was not complete until 2003 (HUD, n.d.b). More recently, HUD introduced additional data quality controls into PIC and, alongside those upgrades, PIC is now known as the Inventory Management System (IMS) of the PIH Information Center, or IMS/PIC.

PHAs electronically transmit data on households enrolled in the PH and HCV programs on either the HUD-50058 or HUD-50058 MTW forms. One should note that HUD did not require the small number (approximately 30) of PHAs participating in the MTW demonstration[12] to submit HUD-50058 data until 2006, and today, MTW agencies have the option of using an abbreviated HUD-50058 MTW form that omits income-based rent calculations.

IMS/PIC includes information on the program in which households participate, and the linked data set condenses some of those codes into larger categories. For the programs included in IMS/PIC, the

[11] MTW PHAs may choose not to include all of their programs within their MTW participation. In general, however, most MTW agencies include their PH and HCV programs (except special purpose vouchers) as part of the MTW program.
[12] For a list of agencies participating in MTW by year, see Webb, Frescoln, and Rohe (2016).

classification scheme is as follows (note that first-order bullets in the list below are the categories included in the linked data set, whereas second-order bullets are categories included in IMS/PIC):

- Public Housing
    - Public Housing
- Housing Choice Vouchers
    - Section 8 Certificate
    - Section 8 Voucher
    - Mod Rehab
    - MTW Tenant-Based Voucher
    - MTW Project-Based Voucher

Tenant Rental Assistance Certification System (TRACS)

For HUD multifamily properties, owners or other responsible entities collect and electronically transmit through the Tenant Rental Assistance Certification System (TRACS) using the HUD-50059 form. Data submitted to TRACS are substantially similar to data transmitted to PIC and include information on family composition, income, date of entry, and reason for the submission. HUD maintains both IMS/PIC and TRACS, rather than a single system, because "recertification rules vary based on programs, and the PHAs that are participating in the MTW demonstration have special rules regarding the timing of recertification" (NCHS, 2016: 51)

Like IMS/PIC, TRACS also includes information on the program in which a household participates. For the programs included in TRACS, the classification scheme is as follows (similar to the previous list, first-order bullets in the list below are the categories included in the linked data set, whereas second-order bullets are categories included in TRACS):

- Project-Based Section 8
    - Project-Based Section 8
- Other Multifamily
    - Section 101 (Rent Supplement)
    - RAP (Rental Assistance Program)
    - Section 236
    - Section 221(d)(3) BMIR (Below Market Interest Rate)
    - Section 202 PRAC (Project Rental Assistance Contract)
    - Section 811 PRAC (Project Rental Assistance Contract)
    - Section 202/162 PAC (Project Assistance Contract)

HUD Data Quality Issues

Although HUD data quality has improved over time, several reports of the U.S. Government Accountability Office (GAO) and HUD's Office of the Inspector General (OIG) have identified HUD data quality issues that were emerging during the early 2000s. As described in one report,

> The primary reason for poor data quality during the early 2000s was due to the migration of data from one system to another…In the early 2000s, data from MTCS were migrated to PIC until the conversion was completed in 2003. This was problematic, however, because prior to the migration, MTCS data was neither cleaned nor validated. Therefore, data migration led to data quality issues. (NCHS, 2016: 17)

Some major errors of the MTCS data, as identified by GAO and OIG reports, included missing addresses, names, or Social Security numbers. Those omissions are considered "fatal" because it is virtually impossible to accurately identify a person without those variables. The fatal errors did not seem to be extremely common, with one report finding that "on average, 7 percent of the data fields contained fatal errors."

Other reports have further found that HUD did not have sufficient data collection or management controls in place during the early 2000s. Specifically, at that time, the PIC system did not ensure that Social Security information was accurately collected. Not only did the system allow for submission of incomplete or inaccurate Social Security numbers, it also did not require use of a common format (e.g., 123-45-6789 vs. 123456789).

For the purposes of this linkage, those data quality limitations are important because they may prevent accurate cross-identification of people who resided in HUD-assisted housing *and* participated in the Add Health study. The risk of incomplete data linkage is especially noteworthy because the period when Add Health participants were adolescents corresponds to the period when HUD data are arguably the least reliable. That noted, HUD undertook several actions to improve data quality during the 2000s, and those efforts have greatly improved the accuracy and completeness of HUD data. For detailed information about these efforts and improving data quality, see Exhibit 1 for a full list of publications on the topic.

Exhibit 1. Publications on HUD Data Quality

Brummet, Quentin. 2014. *Comparison of Survey, Federal, and Commercial Address Data Quality.* Working paper no. 2014-06. Washington, DC: U.S. Census Bureau, Center for Economic Studies. https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-06.pdf.
National Center for Health Statistics (NCHS). 2019. *NCHS-HUD Linked Data: Methodology and Analytic Considerations*. https://www.cdc.gov/nchs/data/datalinkage/NCHS-HUD-Linked-Data-Methodology-and-Analytic-Considerations.pdf.
U.S. Department of Housing and Urban Development (HUD), Office of the Inspector General (OIG). Internal Audit Reports Issued by the Office of the Inspector General 1995–2007. *Review of the Administration of the Portability Features of the Section 8 Housing Choice Voucher Program*. Audit Report No.: 2004-BO-0006. http://archives.hud.gov/offices/oig/reports/oiginter.cfm.
U.S. Department of Housing and Urban Development (HUD), Office of the Inspector General (OIG). Internal Audit Reports Issued by the Office of the Inspector General 1995–2007. *Audit Report on Application Controls over Data Integrity Within the Public and Indian Housing Information Center (PIC)* Audit Report No.: 2004-DP-0003. http://archives.hud.gov/offices/oig/reports/oiginter.cfm.
U.S. Department of Housing and Urban Development (HUD), Office of Policy Development and Research (PD&R). *Quality Control for Rental Subsidy Determination Study*. Various years. https://www.huduser.gov/portal/taxonomy/term/3501.

U.S. Department of Housing and Urban Development (HUD), Office of Public and Indian Housing (PIH). *Notice PIH 2010-3 (HA): Guidance – Verification of Social Security Numbers (SSNs), Social Security (SS) and Supplemental Security Income (SSI) Benefits.* http://portal.hud.gov/hudportal/documents/huddoc?id=DOC_9012.pdf. Published January 2010.

U.S. Government Accountability Office (GAO). *GAO-01-103: Status of Actions to Resolve Serious Internal Control Weaknesses*. http://www.gao.gov/products/GAO-01-103. Published October 16, 2000.

Data Linkage Approval, Eligibility, and Process

Overview

More than 70.6 million HUD individual resident (member) and household-level records were searched to locate the National Longitudinal Study of Adolescent to Adult Health (Add Health) study participants. Add Health and HUD records were linked probabilistically because there was no unique identifier (e.g., Social Security number) available in both data sources. Eight partial identifiers were used to identify probable matches: first, middle, and last name; month, day, and year of birth; sex; and census block group of residence (12-digit Federal Information Processing Standards [FIPS] code). All members of the original Add Health longitudinal cohort (n=20,745) with non-missing data on these eight variables were linkage eligible (see Exhibit 2).

Exhibit 2. Add Health Respondents Eligible for HUD Record Linkage, by Survey Wave and Years

	Wave I	Wave II	Wave III		Wave IV		
	1995	1996	2001	2002	2007	2008	2009
Total Records	20,739	14,738	11,082	4,088	193	15,372	136
Eligible Records	19,806	14,418	10,830	4,059	193	15,367	136
Eligible/Total (%)	95.5%	97.8%	97.7%	99.3%	100.0%	100.0%	100.0%

The initial search of 70.6 million HUD member-level records was restricted to the 7 years in which Add Health was actively surveying participants: 1995 (Wave I), 1996 (Wave II); 2001–2002 (Wave III); and 2007–2009 (Wave IV). Because not all partial identifiers were fixed—particularly residence block group—this approach leveraged the most current (and, hence, the most accurate) partial identifier data on study participants when searching for matches in the vast HUD database.

This initial phase of the linkage, restricted to the seven annual periods, identified 1,159 unique Add Health participants within the HUD records regardless of timing.[13] Next, using the uniquely identifying HUD household head and member identification numbers found from those matches, the search for additional HUD records of the Add Health participants was expanded to the full set of 23 annual HUD administrative files (1995–2017). This expanded search yielded an additional 5,526 non-unique records of Add Health participants (see **Exhibit *8*** for a schematic of the linkage process).

[13] As discussed below, a key variable in the linkage is FIPS codes, for which Add Health has records for respondents only in the year an Add Health survey is completed; thus, the initial linkage was restricted to the seven annual periods in which Add Health was in the field.

Exhibit 3. Number of Add Health Respondents With Non-Missing Data for Each Partial Identifier, by Survey Wave and Years

Partial Identifier	Wave I 1995	Wave II 1996	Wave III 2001	Wave III 2002	Wave IV 2007	Wave IV 2008	Wave IV 2009
Name	19,806	14,418	11,014	4,064	193	15,372	136
Date of Birth	20,736	14,733	11,008	4,087	193	15,370	136
Sex	19,823	14,431	11,014	4,064	193	15,372	136
12-digit FIPS code	20,606	14,655	10,830	4,059	193	15,367	136

Linkage Approval

In compliance with the Federal Policy for the Protection of Human Subjects ("Common Rule"), all Add Health procedures, including the HUD-Add Health Data Linkage Project, were approved by the Institutional Review Board (IRB) at the University of North Carolina (UNC)-Chapel Hill. Written parental or guardian consent and adolescent assent was obtained before the Wave I and II interviews. Similarly, before the Wave III, IV, and V interviews, consent was obtained from all respondents (then adults). As required under 45 CFR 46, all respondents were informed of the study's purpose, expected duration of participation, procedures, possible benefits and risks of participation, procedures for maintaining participant confidentiality, and monetary compensation for participation. Participants were provided with the contact information for the principal investigator and the UNC IRB, and they were informed that their participation was voluntary. Specific consent for linking HUD administrative data to the Add Health database was waived, as provided under 45 CFR 46.116(d), because there was minimal risk to participants, the waiver would not adversely affect the rights and welfare of participants, and the linkage could not be practicably executed without the waiver.

Linkage Eligibility: Add Health

For each of the 7 calendar years (spanning four waves) during which Add Health was in the field actively collecting data, all respondents with non-missing values for the eight partial identifiers were eligible for the data linkage (see Exhibit 2 for specific figures by year of Add Health data collection). Because not all of the partial identifiers (e.g., residential FIPS code) were fixed values over time, this approach was taken to ensure that the most current partial identifiers were utilized in the record linkage.

Exhibit 2 shows that 95.5–100 percent of Add Health records, across all four waves, contained non-missing data on the eight partial identifiers and, consequently, were eligible for linkage with HUD administrative data. Exhibit 3 elaborates on Exhibit 2 by documenting eligible record counts by partial identifier categories.

Linkage Eligibility: HUD

For each of the 23 years of the Add Health project's duration (1995–2017), HUD provided UNC-Chapel Hill's Carolina Population Center (CPC) with an annual household-level file and a corresponding annual household member file, for a total of 46 files. Household files contained variables measuring household characteristics, including household composition, location, and program involvement. Member files

contained household members' personal identifiers—specifically, first, middle, and last name; month, day, and year of birth; and sex.

A unique household ID number included in both the household- and member-level files is used to link the two data files. Data cleaning before the linkage, however, suggested that the household ID was not a unique identifier for a significant fraction of all records. Specifically, two anomalies were detected: (1) records sharing unique member IDs but having different birth dates and (2) unique member IDs that were associated with different household IDs within the same calendar year and form type. The frequency of those two anomalies is documented in Exhibit 4, which indicates that those problems were most common in the 1995 records and decreased significantly over time.[14] All HUD records found to have non-unique household or member IDs were ineligible for linkage (see Exhibit 4 and Exhibit 8: "Internal HUD ID Inconsistency").

Exhibit 4. Percentage of HUD Records With Anomalous Member or Household Identifiers, by Add Health Survey Year

Add Health Survey Year and Wave	HUD Member IDs with Inconsistent Birth Dates	HUD Member IDs with Different Head of Household IDs Within Year and Form Type
1995 (Wave I)	6.6%	23.0%
1996 (Wave II)	5.9%	18.7%
2001 (Wave III)	8.8%	16.0%
2002 (Wave III)	6.9%	15.1%
2007 (Wave IV)	2.1%	9.0%
2008 (Wave IV)	1.2%	5.4%
2009 (Wave IV)	1.3%	5.9%

Linkage Methods

Because there was no unique identifier that would locate Add Health participants in HUD administrative records, a probabilistic-linkage approach was used. Probabilistic linkage attempts to uniquely identify individuals with a set of partial identifiers, such as last name and mother's maiden name (Winkler, 2016). To produce links, probabilistic linkage exploits the discriminatory power of partial identifiers. Specifically, it assigns heavier "agreement weights" to partial identifiers that are particularly unique; for example, an uncommon last name will have more discriminatory power than a common last name and, therefore, would be assigned a stronger agreement weight. After assigning those weights, the overall match or linkage score is calculated as the sum of agreement weights for the full set of partial identifiers. Whereas a low linkage score indicates a low degree of agreement across partial identifiers, a high linkage score indicates the opposite.

[14] Note that the period when these problems are most prevalent corresponds to the period when HUD collected data through MTCS.

Probabilistic linkage also allows for some degree of error or inconsistency across databases in how the partial identifiers are captured (e.g., "Kathy" vs. "Cathy"), yielding higher sensitivity than a deterministic linkage approach that relies on perfect matches. Multiple simulation studies have found that probabilistic linkage consistently outperforms deterministic linkage under a variety of file size, discriminating power, and data quality scenarios (Tromp et al., 2011; Zhu et al., 2015).

The probabilistic linkage was executed using Link Plus (version 2.0), a free and publicly available software program included in the Centers for Disease Control and Prevention's (CDC's) suite of packages developed for use with cancer registry data.[15] Link Plus has been successfully applied to link a variety of databases, including hospital discharge data (Bigback et al., 2015), survey data (Garvey Wilson et al., 2010), and vital records (Zhang et al., 2012).

Add Health staff adhered to the typical steps involved in the linkage process, which included the following:

1. Cleaning data
2. Blocking variable selection
3. Phonetic system specification
4. Match variable selection
5. Match method specification
6. Probability calculation method
7. Cutoff value specification
8. Report generation
9. Manual review

Those steps are described in greater detail in the following paragraphs.

Cleaning Data. As described in the previous section, data cleaning revealed anomalies involving HUD's member IDs and household IDs. Consequently, those anomalous records were dropped before initiating the linkage process. In addition, standard recoding of erroneous or suspicious data occurred. Erroneous or unclear data for sex, date of birth, FIPS code, or name were set to "missing." Out-of-range values on year of birth, based on the Add Health sample's birth years–although not necessarily erroneous—were excluded. Those values were excluded to improve computer performance, as resource limitations were preventing Link Plus from performing linkages on the full data.

Blocking Variable Selection. When attempting to link relatively large databases, evaluating all possible matches (i.e., number of records in database A * number of records in database B) often is computationally impractical. The number of possible matches in the HUD-Add Health annual data sets

[15] Registry Plus, a suite of publicly available software programs for collecting and processing cancer registry data. Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 2020. https://www.cdc.gov/cancer/npcr/.

ranged from approximately 29.7 million pairs (218,678 HUD records * 136 AH records) in 2009 to 10.7 billion pairs (988,322 HUD * 10,830 AH) in 2001. In those circumstances, "blocking" is used to reduce the number of potential matches to those that occur within a subset of data defined by a common characteristic or set of characteristics. The five-digit FIPS code was used as the blocking variable, which limited potential matches between HUD and Add Health records to those individuals residing in the same county.

Phonetic system specification. Phonetic systems code string variables (e.g., names) based on how they are pronounced (e.g., "Cathy" and "Kathy" are treated similarly). The New York State Identification and Intelligence System (NYSIIS), one of the most commonly used systems, was used for the purposes of this linkage.[16] According to the Link Plus documentation, NYSIIS is more distinctive than Soundex, the other commonly used phonetic coding system, providing greater accuracy (including, according to some evidence, with Spanish names).

Match variable selection. As previously identified, eight variables common to both databases were selected as matching variables. Each variable was utilized with a specific matching method to maximize its discriminatory power (see Exhibit 5).

Match method specification. Link Plus offers nine matching methods to fully exploit each type of matching variable employed in the linkage process. For example, use of the "Last Name" match method is recommended if last names are one of the match variables. This matching method utilizes the NYSIIS phonetic code, which accounts for misspellings, hyphenated names, and minor typographical errors. Common names receive relatively low weights whereas rare names are assigned relatively high weights. As shown in Exhibit 5, in addition to the "Last Name" method, we used four other match methods: first name, middle name, date, and exact.

[16] For more information on NYSIIS, see https://naldc.nal.usda.gov/download/27833/PDF.

Exhibit 5. HUD-Add Health Linkage Matching Variables

Data Item	Matching Method	Notes
First Name	First Name	Incorporates both partial (Jaro-Winkler Metric) and value-specific matching and NYSIIS phonetic code. Accounts for minor typographical errors; hyphenated names; misspellings; nicknames. Weights by frequency of name in database.
Middle Name	Middle Name	Accounts for occurrence of middle initial vs. full middle name.
Last Name	Last Name	Incorporates both partial (Jaro-Winkler Metric) and value-specific matching and NYSIIS phonetic code. Accounts for minor typographical errors; hyphenated names; misspellings; nicknames. Weights by frequency of name in database.
Month of Birth	Date	Evaluated as one DOB variable. Incorporates partial matching to account for missing month and/or day values. Checks for month (M) and day (D) swapping and digit transpositions. Weighted by level of agreement (MDY agreement is weighted highest; MY second highest; Y lowest).
Day of Birth	Date	See Month of Birth.
Year of Birth	Date	See Month of Birth.
Sex	Exact	A character-for-character string comparison.
12-Digit FIPS Code	Exact	A character-for-character string comparison.

Probability calculation method. The "Direct Method" (default) was used to derive the M-probabilities. The U-probabilities are calculated from the given data. The M-probability refers to the probability that a matching variable agrees given that the pair under comparison is a true match. The U-probability refers to the probability that a matching variable agrees given that the pair under comparison is not a true match. For example, the probability that month of birth is the same for a non-matching pair is 1/12, or 0.083.[17]

Cutoff value specification. Link Plus generates a linkage score, which provides the overall weight for all matching variables. A higher linkage score indicates a higher likelihood of being a true match. Link Plus recommends a cutoff value setting in the range of 7–10. Following a review of all possible links (using a cutoff of 0, which returns all possible linked records), the team elected to set a cutoff value of 10.

Report generation. Link Plus produces two reports in text format:

- The Non-Match Report: contains records with a linkage score falling below the cutoff value.
- The Linkage Report: contains records with a linkage score equal to or above the cutoff value, reported in descending order by linkage score. The report includes metadata, summary statistics, and the Matching and ID variables for each record pair. The percentage of records

[17] For more detail on M and U probabilities, see Blakely and Salmond (2002).

returning a linked record that were then subject to manual review are as follows (although one should note that these figures include some respondents with more than one match, the best of which was then retained following manual review):

- 1995: 4.1 percent
- 1996: 3.8 percent
- 2001: 1.7 percent
- 2002: 2.9 percent
- 2007: 5.6 percent
- 2008: 1.3 percent
- 2009: 1.0 percent

Manual review. All Link Plus-recommended matches were subjected to further review (see Exhibit 6 for the number of recommended linkages for manual review, by year). First, linked records were evaluated programmatically using SAS software.[18] The match was excluded if (1) the comparison of both first and last names yielded values exceeding 300, as evaluated by SAS's string comparison function COMPGED; (2) date of birth did not match satisfactorily (digit transposition or differences of one were tolerated); or (3) respondent sex disagreed.[19] All remaining records then underwent manual review. This process repeated for each wave of study participation.[20] Where multiple records were submitted for a respondent, the project retained the most reliable of the proposed matches returned by Link Plus and the subsequent vetting process. This process resulted in the rejection of 142 Link Plus-recommended matches (see Exhibit 7).

Exhibit 6. Recommended Linkages Before Manual Review

Corresponding to Add Health Survey Years	Total Unique Respondents Linked	Average Match Score	Minimum Match Score	Maximum Match Score
All	1,337	27.7	10	43.3
1995 (Wave I)	373	25.6	10.1	42.1
1996 (Wave II)	282	27.3	10.1	41.5
2001 (Wave III)	356	29.5	10.2	42.7
2002 (Wave III)	137	26.4	10.2	41.7
2007 (Wave IV)	4	19.4	11.6	24.2
2008 (Wave IV)	551	28.6	10	43.4
2009 (Wave IV)	5	22	12	28.3

Notes: Fewer respondents were linked in 2007 and 2009 given that Add Health fielded far fewer surveys in those years than in 2008 (see exhibit 8). Years shown correspond to Add Health Survey Waves.

[18] SAS Institute Inc., Cary, North Carolina, USA.

[19] When identification of sex was erroneous or unclear, sex entry was set to "missing."

[20] Of note, special attention was paid to ethnic differences in names—especially Hispanic names—when manually reviewing the matches, particularly if last name agreement existed elsewhere in the record—for instance, a parent's last name at some wave.

Exhibit 7. Recommended Linkages After Manual Review

Corresponding to Add Health Survey Years	Total Unique Respondents Linked	Average Match Score	Minimum Match Score	Maximum Match Score
All	1,159	28.6	10.1	43.3

Linkage Results and Summary

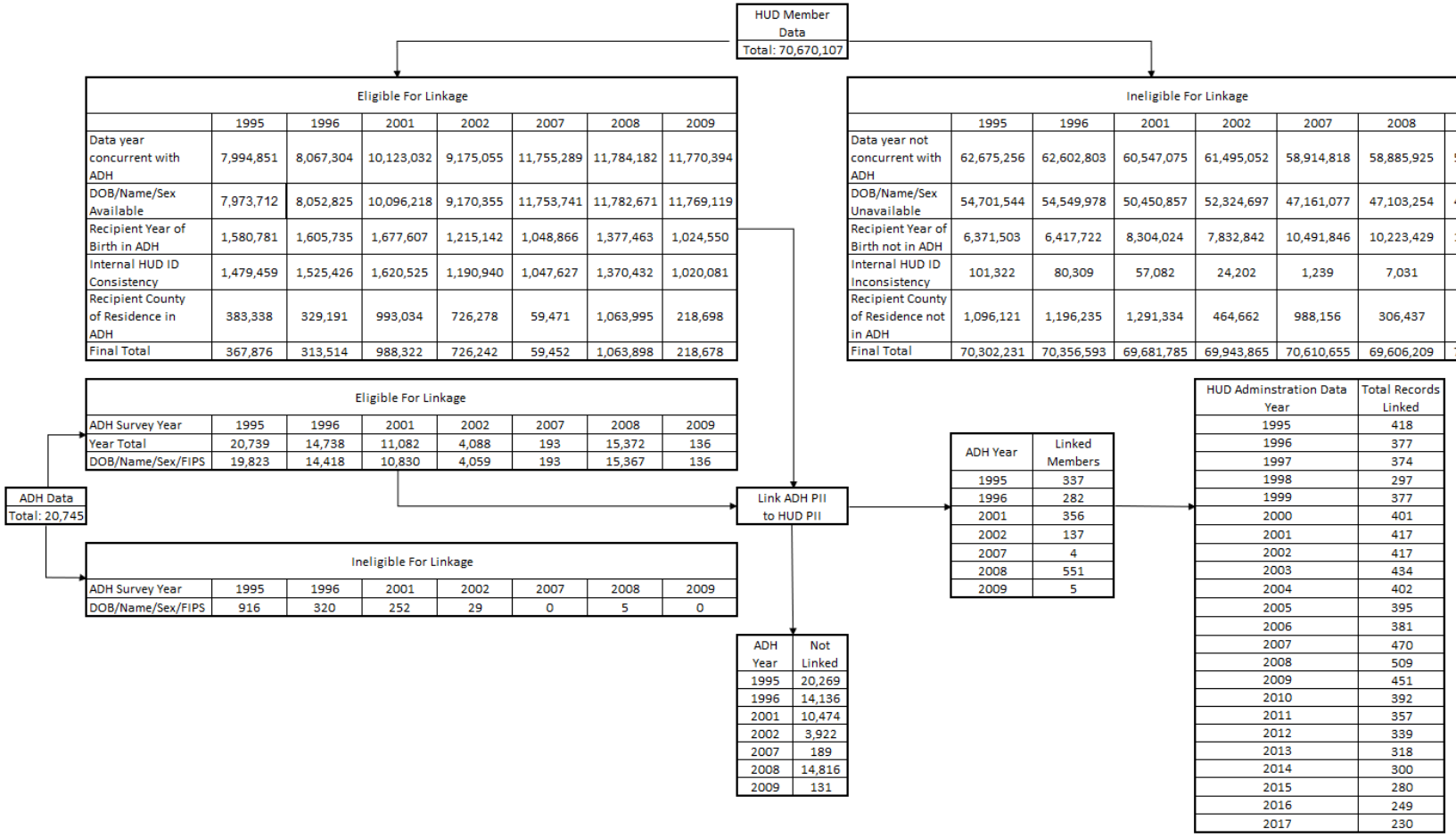
A schematic diagram of the linkage process is presented in Exhibit 8. *Appendix A: Eligible Concurrency Table* contains a more comprehensive table, showing Add Health survey wave year and the sex and age of individuals attached to link-eligible versus ineligible records. The total count of HUD member records (70,670,107) is reported at the top center of the schematic. These HUD records are distributed between eligible and ineligible records in the two large boxes below. Each box contains one column for each of the 7 calendar years that Add Health was actively engaged in data collection. Those 7 years are significant because they are the source of the current name and residential address (12-digit FIPS) of Add Health respondents that are used for matching.

Focusing on the left box, labeled "Eligible for Linkage" and the column for 1995, we see, for example, that approximately 7.99 million of the 70.67 million total HUD member records are for this calendar year. Moving down this column, about 1.58 million records also share the same birth year with an Add Health respondent. The number of eligible HUD records is further reduced to the 383,338 members who also share the same county of residence with Add Health respondents surveyed in 1995.

The second box on the left side of Exhibit 8 reports the counts of eligible Add Health (ADH) respondents from each calendar year, and the third box reports the counts of ineligible Add Health respondents. Add Health eligibility was based on having complete data on the partial (quasi) identifiers.

Finally, moving to the bottom right side of Exhibit 8, we see the linkage counts, by Add Health data collection year. The box in the bottom right shows that, following completion of the linkage process, HUD member IDs that were successfully attached to Add Health respondents within a given survey year were used to identify records belonging to those respondents outside those survey years. This process required the retention of cases with head-of-household IDs associated with member IDs that had been linked to Add Health respondents within each survey year. Any record from 1995 to 2017 with those head-of-household IDs was then retained in the final linked data set.

Exhibit 8. HUD-Add Health Data Linkage Schematic



Eligible For Linkage	1995	1996	2001	2002	2007	2008	2009
Data year concurrent with ADH	7,994,851	8,067,304	10,123,032	9,175,055	11,755,289	11,784,182	11,770,394
DOB/Name/Sex Available	7,973,712	8,052,825	10,096,218	9,170,355	11,753,741	11,782,671	11,769,119
Recipient Year of Birth in ADH	1,580,781	1,605,735	1,677,607	1,215,142	1,048,866	1,377,463	1,024,550
Internal HUD ID Consistency	1,479,459	1,525,426	1,620,525	1,190,940	1,047,627	1,370,432	1,020,081
Recipient County of Residence in ADH	383,338	329,191	993,034	726,278	59,471	1,063,995	218,698
Final Total	367,876	313,514	988,322	726,242	59,452	1,063,898	218,678

Ineligible For Linkage	1995	1996	2001	2002	2007	2008	2009
Data year not concurrent with ADH	62,675,256	62,602,803	60,547,075	61,495,052	58,914,818	58,885,925	58,899,713
DOB/Name/Sex Unavailable	54,701,544	54,549,978	50,450,857	52,324,697	47,161,077	47,103,254	47,130,594
Recipient Year of Birth not in ADH	6,371,503	6,417,722	8,304,024	7,832,842	10,491,846	10,223,429	10,531,155
Internal HUD ID Inconsistency	101,322	80,309	57,082	24,202	1,239	7,031	4,469
Recipient County of Residence not in ADH	1,096,121	1,196,235	1,291,334	464,662	988,156	306,437	801,383
Final Total	70,302,231	70,356,593	69,681,785	69,943,865	70,610,655	69,606,209	70,451,429

Eligible For Linkage	1995	1996	2001	2002	2007	2008	2009
ADH Survey Year							
Year Total	20,739	14,738	11,082	4,088	193	15,372	136
DOB/Name/Sex/FIPS	19,823	14,418	10,830	4,059	193	15,367	136

Ineligible For Linkage	1995	1996	2001	2002	2007	2008	2009
ADH Survey Year							
DOB/Name/Sex/FIPS	916	320	252	29	0	5	0

ADH Year	Not Linked
1995	20,269
1996	14,136
2001	10,474
2002	3,922
2007	189
2008	14,816
2009	131

ADH Year	Linked Members
1995	337
1996	282
2001	356
2002	137
2007	4
2008	551
2009	5

HUD Adminstration Data Year	Total Records Linked
1995	418
1996	377
1997	374
1998	297
1999	377
2000	401
2001	417
2002	417
2003	434
2004	402
2005	395
2006	381
2007	470
2008	509
2009	451
2010	392
2011	357
2012	339
2013	318
2014	300
2015	280
2016	249
2017	230

Description of Linked Data Files

Program and household variables of interest were retained, and Add Health's disseminated unique respondent ID (AID) replaced the temporary internal project ID. This resulted in a hierarchical file (i.e., long-format data file) by respondent ID, data year, program type, and activity, containing a total of 8,585 records, representing 1,159 Add Health sample members. Variables of interest from the HUD database appearing in the Adverse Childhood Experience data file include the following (see data codebook from Add Health's website for full description – https://addhealth.cpc.unc.edu/wp-content/uploads/docs/restricted_use/HUD-Assisted_Housing_Supplementary_Data.zip):

- AID—Add Health respondent identifier
- ACE01Y—HUD-assisted housing record year
- ACE01Q—HUD-assisted housing record quarter
- ACE02—Form type for assisted housing record
- ACE03—Program type
    - PH
    - HCV
    - MF: Project-Based Section 8
    - MF: Other multifamily
- ACE04—Transaction record type
- ACE05—Total household members
- ACE06—Episode demarcation flag

The final chapter of this document discusses the storage of the linked data set and how researchers can obtain access to it.

Summary Statistics on Linked HUD-Add Health Data

Using the hierarchical data file created by the linkage process, this section outlines summary and descriptive statistics on linked sample members' housing episodes and concurrency between HUD-assisted housing episodes and participation in the National Longitudinal Study of Adolescent to Adult Health (Add Health) surveys. Although HUD-assisted housing episodes were calculated by specific date, the resulting output can be summarized only at the quarter-year level due to Add Health data confidentiality requirements.

This section briefly describes the methodology used to calculate episodes and provides descriptive statistics about those episodes. Overall, we find that most linked sample members (82 percent) experienced four or fewer episodes across the study period and that the average length of their housing episodes were quite short, at just 1.5 years. In the next section, we use the housing episodes variable to calculate concurrency between HUD-assisted housing residence and Add Health survey participation.

Episodes

Methodology

The term "episodes" refers to the specific time periods when linked sample members lived in HUD-assisted housing. Episodes were calculated on the basis of HUD transaction records (see discussion earlier regarding when a transaction occurs). One challenge of calculating episodes on the basis of those records is determining whether "gaps" between HUD transactions represent a single episode or two distinct housing episodes. For example, following program entry, transactions occur only annually for most programs—provided the program is not a Moving to Work (MTW) public housing authority (PHA)—thus, gaps of a year or more in HUD transaction records often occur.

Further, PHAs participating in MTW do not have to recertify clients as frequently as non-MTW sites; often, they only have to recertify biennially rather than annually. Thus, a long break in linked sample members' transaction records may simply reflect that they live at an MTW site, not necessarily that they experienced two distinct HUD-assisted housing episodes. Further, even PHAs that are required to recertify clients on an annual basis are given "60 days' leeway to submit reports"; therefore, in some instances, longer breaks simply reflect a PHA taking advantage of that leeway (NCHS, 2019: 28).

Fortunately, researchers developed clear standards for calculating HUD-assisted housing episodes when completing the National Center for Health Statistics(NCHS)-HUD data linkage (NCHS, 2019: 28). Those same standards were used to calculate episodes for this linkage:

- Non-MTW records: if a break in transaction records was 425 days or more, those records were to represent two distinct episodes. If the break was shorter, those records were considered to represent a single episode. The 425-day period is the standard because non-MTW sites must complete recertifications every 425 days (1 year plus 60 days of leeway).

- MTW records: if a break in transaction records was 790 days or more, those records were considered to represent two distinct episodes. If the break was shorter, those records were considered to represent a single episode. The 790-day period is used as the threshold because most MTW sites must complete recertifications every 790 days.

On the basis of those criteria, the research team created a housing episode "flag"—that is, a variable that allows researchers to determine when one housing episode ends and a new one begins—for each transaction record included in the linked sample. This flag is a three-digit code that adheres to the format outlined in Exhibit 9. Included in appendix B is R code that will allow researchers to format the linked, hierarchical data file to compute the summary statistics on both housing episodes and concurrency.

Exhibit 9. Flag for Housing Episode in Linked Data Set

Code	MTW Site?	Description	% of All Linked Transaction Records
100	Yes	Beginning of first episode	0.61
101	Yes	Beginning of a subsequent episode	0.04
102	Yes	Continuation of an episode	2.31
200	No	Beginning of first episode	13.31
201	No	Beginning of a subsequent episode	24.69
202	No	Continuation of an episode	59.04

Summary Statistics

Most linked sample members (82.1 percent) experienced four or fewer HUD-assisted housing episodes (see Exhibit 10); in fact, nearly one-half (49.7 percent) of all respondents experienced only one episode. The data have a long tail, though, and some respondents experienced up to eight unique episodes in HUD-assisted housing.

Exhibit 10. Number of HUD-Assisted Housing Episodes by Linked Sample Members

# of Unique Episodes	N	%	Cumulative %
1	576	49.7	49.7
2	329	28.4	78.1
3	149	12.9	90.9
4	64	5.5	96.5
5	26	2.2	98.7
6	11	0.9	99.7
7	*	*	*
8	*	*	*
Total	1,159	100%	100%

*-Suppressed

Exhibit 11 provides summary statistics for all linked respondents' episodes by episode number. This exhibit can be interpreted as follows: the average length of a linked respondent's first episode was 3.21 years; of linked respondents who had a second episode, the average length was 2.99 years; and so on. As this exhibit shows, the average length of linked respondents' housing episodes is relatively consistent across each episode—with the exception of the small number of respondents who had a seventh or eighth episode—at approximately 2.2 to 3.2 years. The standard deviation for each of these episode numbers is somewhat large relative to the mean, however, which indicates significant variability in the length of those episodes.

Exhibit 11. Episode Length (Years) by Episode Number

Episode	Min	25th Percentile	Mean	75th Percentile	Max	SD	N
1	0.25	0.25	3.21	4.25	22.30	4.09	1,159
2	0.25	0.25	2.99	4.25	19.70	3.58	583
3	0.25	0.25	2.75	3.75	17.00	3.32	254
4	0.25	0.25	3.04	4.00	13.70	3.60	105
5	0.25	0.25	2.16	2.50	13.20	3.25	41
6	0.25	0.50	2.98	4.12	9.00	2.83	15
7	*	*	*	*	*	*	*
8	*	*	*	*	*	*	*

*-Suppressed

Exhibit 12 breaks down episodes by program. This exhibit can be interpreted as follows: all public housing episodes—regardless of the number of episodes—are listed in the "Public Housing" column. Across the HUD program types, average episodes in the public housing program are longest (mean 4.36 years), followed by the HCV program (mean 3.89 years). Average episodes in the Project-Based Section 8 program (1.55 years) and the Other Multifamily program (1.29 years) are considerably shorter in comparison.

Exhibit 12. Episode Characteristics (Years) by Program Type

Program Category	Min	25th Percentile	Mean	75th Percentile	Max	SD	Number of Episodes
Public Housing	0.25	1.00	4.36	6.75	22.30	2.02	391
HCV	0.25	0.75	3.89	5.75	22.00	1.94	935
Project-Based Section 8	0.25	0.25	1.55	2.00	20.00	2.19	745
Other Multifamily	0.25	0.25	1.29	1.13	12.30	2.43	91

Exhibit 13 provides information on which HUD-assisted programs' linked sample members participated. Nearly three-fourths (73 percent) of linked sample members have episodes in only one HUD-assisted program, with a plurality (42 percent) of linked sample members participating in only the Housing Choice Voucher (HCV) program. Fewer linked sample members participated in only the Project-Based Section 8 or public housing (PH) programs (18 and 10 percent, respectively). More than one-fourth of linked sample members participated in two or more HUD-assisted programs during the study period.

Considering the time at which our data were collected and the household structure (i.e., young families) of linked sample members, those episode length statistics are comparable to episode lengths identified in previous studies. For example, McClure's (2018) analysis of length of stay in HUD-assisted housing found that from 1995 through 2015, the average length of stay for non-senior families with children ranged between 1.9 and 4.9 years, which is roughly the range of the mean episode lengths identified in this study.

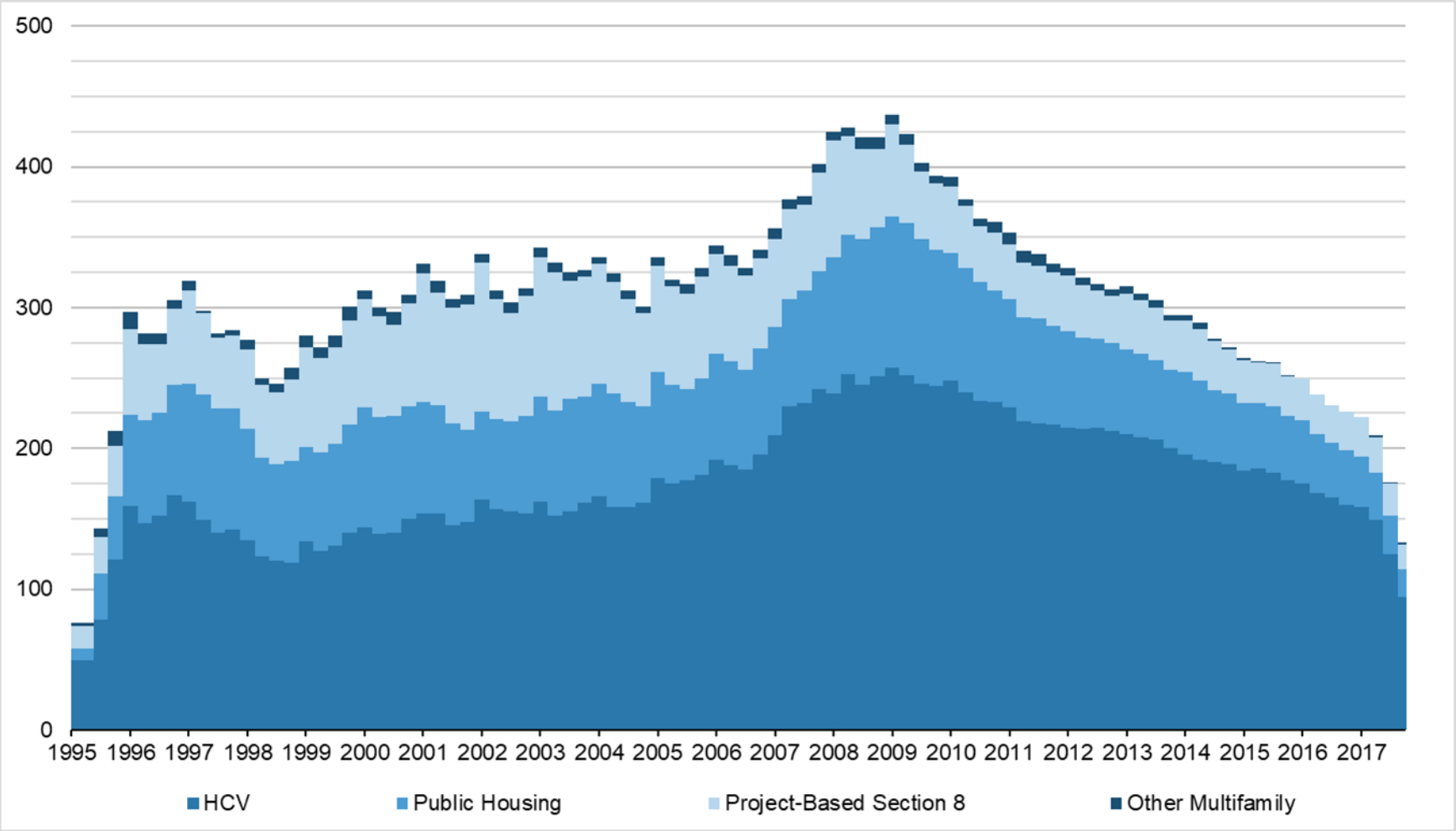
Exhibit 13. HUD-Assisted Programs in Which Linked Sample Members Participated

HUD-assisted program	N	%
One program only	*851*	*73%*
HCV	488	42%
Project-based Section 8	206	18%
Public housing	126	10%
Other multifamily	31	3%
Two or more programs	308	27%
Total	**1,159**	**100%**

Exhibit 14 provides further information on episodes, by program, by charting the number of active HUD-assisted housing episodes by quarter (Q). From Q1 1996 to Q2 2007, the number of active episodes consistently ranged between 275 and 325 (except for a brief dip in 1998 and 1999). After that time, the number of active episodes rose rapidly, reaching a high of roughly 425 linked sample members in HUD-assisted housing from mid-2007 through mid-2009. After mid-2009, the sample size gradually decreased.

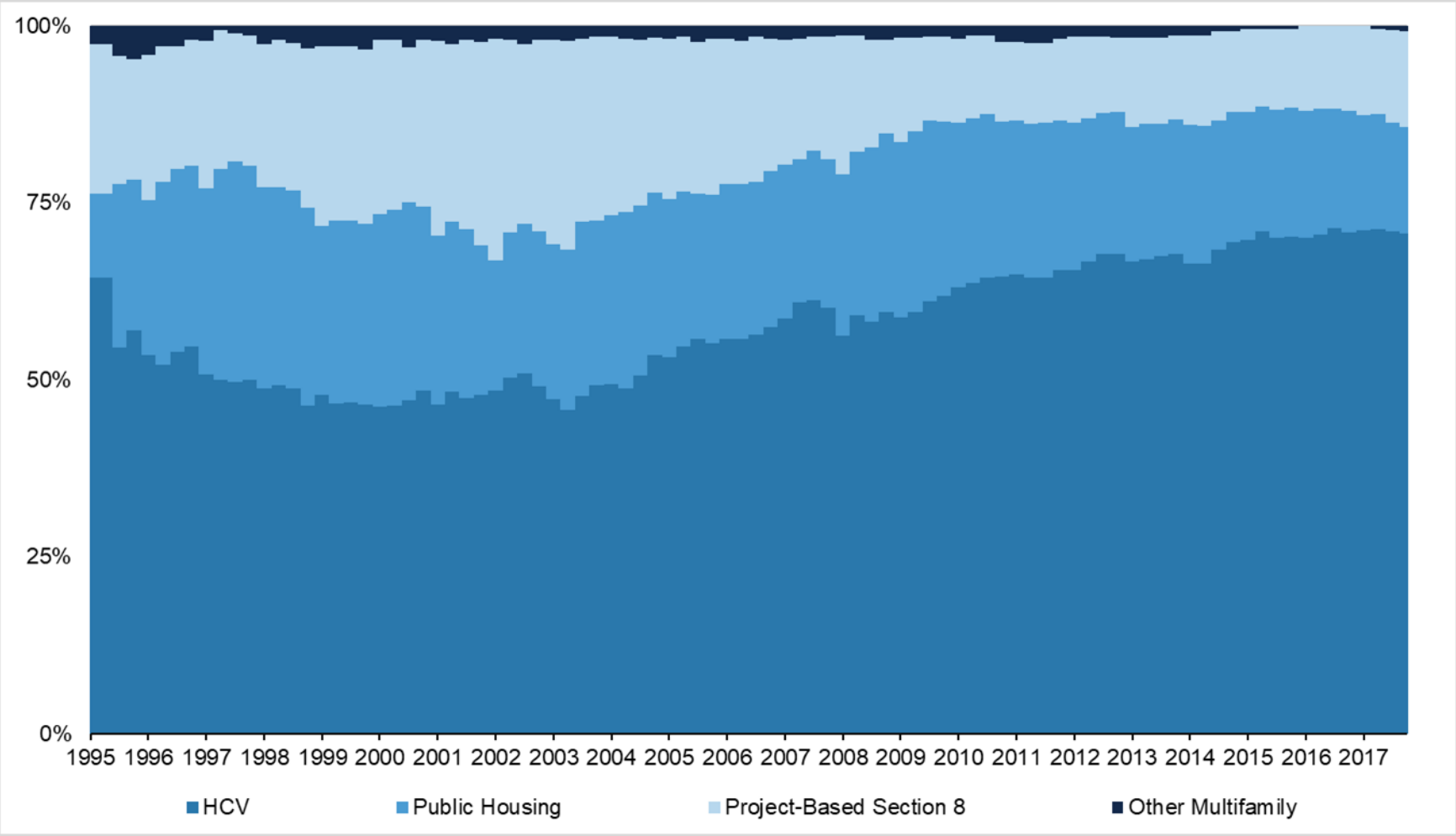
Exhibit 15 shows that, in every quarter, participants in the HCV program made up a plurality (and, in many quarters, a majority) of linked sample members. Furthermore, the proportion of linked sample members in the HCV program increased across the period charted; beginning in roughly 2006, HCV participants constituted most linked sample members.

Exhibit 14. Linked Sample Members in HUD-Assisted Housing by Program Type by Quarter



HCV = Housing Choice Voucher.

Exhibit 15. Proportion of Linked Sample Members in HUD-Assisted Housing Programs by Quarter



HCV = Housing Choice Voucher.

Concurrency

Methodology

Although many linked sample members had active HUD-assisted housing episodes during a given wave of the Add Health survey, not all those linked sample members responded to the survey. Researchers should thus determine whether "concurrency" exists with respect to linked sample members' housing episodes and participation in Add Health surveys.

For the purposes of this report, concurrent episodes were calculated as follows: if a linked sample member (1) had an active episode during an Add Health survey wave and (2) responded to that survey, then that episode was counted as "concurrent." Episodes that met only the first of those conditions was considered to be a "nonrespondent" episode, as those linked sample members were living in HUD-assisted housing at the time of the survey but did not respond. Episodes that met neither of those conditions were considered "ineligible" respondents and were excluded from the summary statistics in this section. Note the absence of "nonresponse" episodes at Wave I because all linked sample members responded to that survey.

Exhibit 16 provides an infographic on the procedure that was used to identify whether a given episode could be considered concurrent or not. Whereas "X" indicates the years that a person was living in HUD-assisted housing, green shading indicates the quarter when that person responded to the Add Health survey. Although this infographic specifically outlines how concurrency was calculated at Wave III of Add Health data collection, which occurred between 2001 and 2002, it could apply to any wave of data collection, as our procedure was identical across all waves.

Exhibit 16. Example of Add Health-HUD Concurrency

	Wave III Data Collection Period											
	2001				2002				2003			
Designations	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1. Concurrent	X	X	X	X	X	X	X	X	X	X	X	X
2. Nonresponse				X	X	X	X	X	X	X		

Q = quarter.

Notes: The X's represent residence in HUD-assisted housing in a given quarter, and green-shaded cells represent if—and if so, when—a respondent completed an Add Health survey. **Concurrent:** Lived in HUD housing in the same quarter as Add Health data collection and ***responded*** to the survey. **Nonrespondent:** Lived in HUD housing in the same quarter as Add Health data collection but ***did not respond*** to the survey.

Descriptive Characteristics

Exhibit 17 shows the number of linked sample members with a concurrent versus nonresponse episode for each wave of Add Health data collection. During that period, the number of concurrent observations increased across the waves: whereas Wave I yielded only 182 concurrent observations, Wave III and IV yielded roughly 300 and 400 concurrent observations, respectively. That increase likely occurred because

the overall number of linked sample members living in HUD-assisted housing increased between Waves I and III (1995–1996), respectively, and Waves III and IV (2001–2002 and 2008–2009) (**Exhibit *14***).

Across each wave, stable trends emerged with respect to program type and race among concurrent and nonresponse linked sample members. Regarding program type, at each wave, a plurality of concurrent and nonresponse linked members were participating in the HCV program, followed by PH, Project-Based Section 8, and other multifamily (MF) programs. A stable racial composition exists across each, with non-Hispanic African Americans[21] constituting the majority of both concurrent and nonresponse linked sample members, followed by non-Hispanic Whites, Hispanics or Latinos (of any race), and other/multiracial.

By contrast, a substantial change in gender composition occurred across waves. At Waves I and II, slightly less than 40 percent of respondents were males, and the remaining 60 percent were females. At Waves III and IV, however, males accounted for less than 25 percent of all concurrent and nonresponse linked sample members, whereas females accounted for 75 percent or more. This increase likely occurred because many PHAs have preferential admission policies for parents with children who are at risk for homelessness—many of whom are females—therefore, female heads of household are generally overrepresented in HUD-assisted housing (Eggers, 2017; USICH, 2013). As linked sample members begin to age into adulthood and have children, therefore, females are overrepresented in the concurrent and nonresponse categories.

Exhibit 18 shows the number of concurrent Add Health survey responses across all linked sample members. Again, concurrency here is measured by quarter due to Add Health data restrictions. A plurality of linked sample members is concurrent for a single Add Health wave—meaning they completed an Add Health survey in the same quarter they resided in HUD-assisted housing. Smaller numbers of linked sample members were concurrent for two or more waves. One should note that a substantial number (approximately 30 percent) of linked sample members never resided in HUD-assisted housing in the same quarter as they completed an Add Health survey.

[21] One should note that the Add Health survey instrument utilized "Black" and not "African-American" and thus this category includes some foreign-born Blacks (approximately 3 percent of the Add Health sample members who identified as "Black") who may not identify as African-American.

Exhibit 17. HUD/Add Health Concurrency by Add Health Wave

	Wave I (1995)	Wave II (1996)		Wave III (2001–2002)		Wave IV (2008)		Total (1995–2008)	
	Concurrent	Concurrent	Non response	Concurrent	Non response	Concurrent	Non response	Concurrent	Non response
Program Type									
Public Housing	42	59	79	58	73	102	113	261	265
HCV	105	134	169	146	183	229	303	614	655
Project-Based Section 8	*	*	*	*	130	*	88	221	273
Other Multifamily	*	*	*	*	13	*	12	22	31
Gender									
Males	74	97	121	57	84	50	79	278	284
Females	108	143	188	242	315	347	437	840	940
Race									
Non-Hispanic White	51	64	73	68	92	91	123	274	288
Non-Hispanic African American	103	135	188	189	256	273	340	700	784
Hispanic or Latino (any race)	*	*	*	*	*	*	*	134	*
Other/Multiracial	*	*	*	*	*	*	*	10	*
Total Linked Sample Members	*182*	*240*	*309*	*299*	*399*	*397*	*516*	*1,118*	*1,224*

* - Suppressed

Concurrent: Resided in HUD housing same quarter as Add Health data collection and ***responded*** to Add Health survey.
Nonresponse: Resided in HUD housing during survey Add Health data collection but ***did not respond*** to Add Health survey.

Exhibit 18. Number of Concurrent Add Health Survey Responses by Linked Sample Members.

		Concurrent				
	Never Concurrent	Single Wave	Two Waves	Three Waves	Four Waves	Sum
Program Type						
Public Housing	38	117	47	*	*	226
HCV	155	297	93	*	*	568
Project-Based Section 8	115	158	*	*	*	318
Other Multifamily	29	14	*	*	*	47
Gender						
Males	112	132	51	*	*	309
Females	225	454	131	*	*	850
Race						
Non-Hispanic White	107	169	37	*	*	323
Non-Hispanic African-American	179	347	117	*	*	681
Hispanic or Latino (any race)	*	*	*	*	*	*
Other/Multiracial	*	*	*	*	*	*
Total Linked Sample Members	*337*	*586*	*182*	*	*	*1,159*

**-Suppressed*

Note: "Never Concurrent" are still included in the linked data set.

Linked Data Storage and Access

Although the linked HUD-National Longitudinal Study of Adolescent to Adult Health (Add Health) data set excludes all personal identifiers, and the data have undergone deductive or statistical disclosure risk assessment, they are considered restricted-access data and will not be released as public-use files. Consequently, researchers interested in obtaining access to the linked data set, as well as any other restricted-use Add Health data, must apply for a restricted-use contract using the Carolina Population Center (CPC) Data Portal.[22] The restricted-use contract is a data-use agreement between the University of North Carolina (UNC)-Chapel Hill and the institution that is requesting the data file. This agreement is signed by an institutional representative of each institution.

Key elements of the restricted-use contract include the following:[23]

- Completion of the Data Use Agreement (DUA), in which the applicant must provide information about the project, agree to CPC's terms for usage of restricted-use data, and secure the signature of an institutional representative.
- A security plan for the sensitive data (see below).
- Data justification, in which the applicant states how the requested data sets relate to his or her research.
- Supplemental agreements signed by all other researchers (those with authorization to access Add Health data).
- Security pledges signed by the principal investigator, other researchers, collaborators (those involved in the research but who do not access the Add Health data), information technology (IT) staff, and officemates of the principal investigator, co-investigators, and research staff.
- Payment of the restricted-use data set fee (as of 2020, the fee is $1,000 for a new restricted-use data contract; some Add Health data sets require an additional fee).
- Institutional Review Board approval letter for the research project.
- A designated downloader form.

Regarding the sensitive data security plan, Add Health allows for restricted-use data to be stored on an encrypted stand-alone desktop computer, an encrypted external hard drive, or a secure server. Regarding the latter, the server may either be a compute server (for which files are stored on the server, and the processing of all data is done on the server) or a file server (for which files are stored on the server, but processing occurs on a user's computer). Regardless of storage option, applicants must complete and have approved a data security plan before receiving restricted-use data.

Furthermore, Add Health Restricted-Use Data Contracts require that investigators submit annual reports to the Add Health Contract Administrator. The annual report must include a current IRB letter; list

[22] The CPC data portal is accessed online at https://data.cpc.unc.edu/.
[23] All information in this section is current as of September 2020 but is superseded by any information on the CPC or Add Health websites or in written communication from any CPC or Add Health staff member.

of public presentations and papers accepted for publication, using results based on Add Health data; grants awarded for use of Add Health data; and dissertations or theses completed using Add Health data. In addition, investigators must provide a list of those associated with the contract (researchers, collaborators, administrative support, IT staff, and officemates) and a list of individuals previously associated with the contract but who are no longer associated with it.

Appendix A: Eligible Concurrency Table

The following table describes how many HUD-member records were eligible for linkage for each year of the National Longitudinal Study of Adolescent to Adult Health (Add Health) data collection during the study period and provides sex and age composition of those member records. As described in the section titled “Linkage Eligibility: HUD,” HUD-member records were considered eligible for linkage only if they had the required partial identifier information and a reliable household/member ID. Furthermore, given the ages of Add Health participants, the “HUD Eligible for Linkage” column is limited to those in the age range of Add Health sample members.

Exhibit 19. HUD-Add Health Concurrency

			Total in HUD Data	HUD Eligible for Linkage	HUD-Add Health Linked
1995 (Wave I)	Sex	Male	2,869,389	173,988	160
		Female	5,104,323	208,757	249
	Age	0–17	3,179,724	253,437	343
		18–39	2,345,313	129,308	66
		40–64	1,195,043		
		65–120	1,253,632		
1996 (Wave II)	Sex	Male	2,926,336	146,814	122
		Female	5,126,489	182,469	178
	Age	0–17	3,349,244	199,043	233
		18–39	2,261,472	130,240	67
		40–64	1,193,889		
		65–120	1,248,220		
2001 (Wave III)	Sex	Male	3,690,567	320,563	76
		Female	6,405,651	679,635	292
	Age	0–17	4,309,831	116,724	
		18–39	2,690,494	883,474	368
		40–64	172,677		
		65–120	136,916		
2002 (Wave III)	Sex	Male	3,350,565	193,898	63
		Female	5,819,790	542,868	303
	Age	0–17	3,897,162		
		18–39	2,415,472	736,766	366
		40–64	1,583,613		
		65–120	1,274,108		
2007 (Wave IV)	Sex	Male	4,324,303	7,510	50
		Female	7,429,438	53,268	367
	Age	0–17	4,847,246		
		18–39	3,040,418	60,778	417
		40–64	2,281,351		
		65–120	1,584,726		
2008 (Wave IV)	Sex	Male	4,335,337	197,481	60
		Female	7,447,334	884,241	384
	Age	0–17	4,813,591		
		18–39	3,055,238	1,081,722	444
		40–64	2,319,222		
		65–120	1,594,620		
2009 (Wave IV)	Sex	Male	4,344,528	45,345	55
		Female	7,424,591	177,938	343
	Age	0–17	4,722,167		
		18–39	3,071,693	223,283	398
		40–64	2,367,517		
		65–120	1,607,742		

Appendix B: R Code for Concurrency and Episode Calculations

```
# Project: HUD-Add Health Episodes & Concurrency
# Author: Atticus Jaramillo, atticusa@live.unc.edu
# Updated: 2/10/2021

library(dplyr)
library(Hmisc)
library(tidyr)
library(reshape2)
library(lubridate)

########################################
## Load Data & update file paths as necessary ##
########################################

HUD<-sasxport.get("/ifs/sec/cpc/addhealth/users/atticusa/ACEW1234/ACEW1234.xpt")
wave1<-sasxport.get("/ifs/sec/cpc/addhealth/addhealthdata/wave1/allwave1.xpt")
wave2<-sasxport.get("/ifs/sec/cpc/addhealth/addhealthdata/wave2/wave2.xpt")
wave3<-sasxport.get("/ifs/sec/cpc/addhealth/addhealthdata/wave3/wave3.xpt")
wave4<-sasxport.get("/ifs/sec/cpc/addhealth/addhealthdata/wave4/wave4.xpt")
w1_interview<-select(wave1,aid,imonth,iday,iyear)
w2_interview<-select(wave2,aid,imonth2,iday2,iyear2)
w3_interview<-select(wave3,aid,imonth3,iday3,iyear3)
w4_interview<-select(wave4,aid,imonth4,iday4,iyear4)
demographics<-select(wave1,aid,bio.sex,h1gi1m,h1gi1y,h1gi3,h1gi4,h1gi6a,h1gi6b,h1gi6c,
                     h1gi6d,h1gi6e)
colnames(demographics)<-c("aid","sex","b_month","b_year","age_move_current","latino",
                          "white","black","native_american","asian","other")

##################
## Format variables##
##################
HUD$aid<-as.character(HUD$aid)
HUD$ace01y<-as.numeric(HUD$ace01y)
HUD$ace01q<-as.numeric(HUD$ace01q)
HUD$ace02<-as.numeric(HUD$ace02)
HUD$ace03<-as.numeric(HUD$ace03)
HUD$ace04<-as.numeric(HUD$ace04)
HUD$ace05<-as.numeric(HUD$ace05)
HUD$ace06<-as.numeric(HUD$ace06)

################################
## Identify beginning of all Episodes ##
################################

HUD$start<-
ifelse(HUD$ace06=="100"|HUD$ace06=="200"|HUD$ace06=="101"|HUD$ace06=="201",1,0)

#####################################################
## Format HUD Record Dates to Determine Correspondence ##
```

```
#####################################################

HUD<-HUD %>% group_by(aid) %>%
  arrange(ace01y,ace01q) %>%
  mutate(episode=ifelse(start=="1",0+cumsum(start),start))

HUD<-HUD %>%
  mutate(year_quarter=paste(ace01y,"-",ace01q,sep = ""))

HUD$year_quarter <- gsub("-1", "-01", HUD$year_quarter)       # map first quarter to January
HUD$year_quarter <- gsub("-2", "-04", HUD$year_quarter)       # map second quarter to April
HUD$year_quarter <- gsub("-3", "-07", HUD$year_quarter)       # map third quarter to July
HUD$year_quarter <- gsub("-4", "-10", HUD$year_quarter)       # map fourth quarter October

HUD$year_quarter <- paste(HUD$year_quarter, "-01", sep="")    # add first day of the month

HUD$rec_date <- as.Date(HUD$year_quarter, "%Y-%m-%d")

#################################################
## Fill in dates between each record per-respondent ##
#################################################

HUD<-HUD%>%
  mutate(episode=na_if(episode,"0")) #replace zeros so fill command will work; only works on NA's

HUD<-HUD %>%
  arrange(rec_date,desc(aid))%>%
  fill(episode)
#warnings indicate lost "labels" from data titles, but no impact to underlying data.
# To reapply labels if desired see: https://github.com/amices/mice/issues/157

HUD_expanded<- HUD %>%
  group_by(aid,episode) %>%
  complete(rec_date = seq.Date(min(rec_date), max(rec_date), by="quarter"))

HUD_expanded<- HUD_expanded %>%
  arrange(rec_date,desc(aid)) %>%
  fill(ace03)

HUD_expanded<-select(HUD_expanded,aid,rec_date,episode,ace03)

#################################################
## Merge Add Health Interview Date & Demographics##
#################################################

HUD_expanded<-merge(HUD_expanded,w1_interview,by="aid",all.x=TRUE)
HUD_expanded<-merge(HUD_expanded,demographics,by="aid",all.x=TRUE)

# Keep all linked records (i.e., “all.x=TRUE” to facilitate analysis of non-response)

#########################################
```

```
## Format Add Health Interview Date String ##
########################################

HUD_expanded$imonth<-sprintf("%02d",HUD_expanded$imonth)
HUD_expanded$iday<-sprintf("%02d",HUD_expanded$iday)

HUD_expanded <-HUD_expanded %>%
  mutate(iyear_quarter=paste("19",iyear,"-",imonth,"-",iday,sep = ""))

HUD_expanded$iDate1 <- as.Date(HUD_expanded$iyear_quarter, "%Y-%m-%d")

replace<-as.Date("1995-04-01") ## Add Health did not begin data collection until Q2 1995. Therefore, HUD transaction records from Q1 1995 must be recoded so they are not arbitrarily dropped from correspondence calculations.

HUD <- HUD %>%
  mutate(rec_date=if_else(rec_date=="1995-01-01",replace,rec_date))

HUD_expanded$iDate_correspond1<-floor_date(HUD_expanded$iDate1,"quarters")

HUD_expanded<- HUD_expanded %>%
  mutate(w1_correspond=if_else(rec_date==iDate_correspond1,1,0))

w1_correspond <- filter(HUD_expanded,w1_correspond=="1")
w1_correspond$delete<-duplicated(w1_correspond$aid)
w1_correspond<-filter(w1_correspond,delete==FALSE)

#################################################
## Wave 2 (repeat same process + non-response analysis) ##
#################################################

HUD_expanded<-merge(HUD_expanded,w2_interview,by="aid",all.x = TRUE)

HUD_expanded$imonth2<-sprintf("%02d",HUD_expanded$imonth2)
HUD_expanded$iday2<-sprintf("%02d",HUD_expanded$iday2)

HUD_expanded<-HUD_expanded %>%
  mutate(iyear_quarter2=paste("19",iyear2,"-",imonth2,"-",iday2,sep = ""))

HUD_expanded$iDate2 <- as.Date(HUD_expanded$iyear_quarter2, "%Y-%m-%d")

HUD_expanded$iDate_correspond2 <- floor_date(HUD_expanded$iDate2,"quarter")

HUD_expanded<-HUD_expanded %>%
  mutate(w2_correspond=ifelse(rec_date==iDate_correspond2,1,0))

w2_correspond<-filter(HUD_expanded,w2_correspond=="1")
w2_correspond$delete<-duplicated(w2_correspond$aid)
w2_correspond<-filter(w2_correspond,delete==FALSE)

HUD_expanded<-HUD_expanded %>%
```

```
  mutate(non_resp2=if_else(is.na(iDate_correspond2) &
                           rec_date=="1996-04-01" |
                           rec_date=="1996-07-01",1,0))

w2_nonresponse<-filter(HUD_expanded,non_resp2=="1")
w2_nonresponse$delete<-duplicated(w2_nonresponse$aid)
w2_nonresponse<-filter(w2_nonresponse,delete==FALSE)

###########
## Wave 3 ##
###########

HUD_expanded<-merge(HUD_expanded,w3_interview,by="aid",all.x = TRUE)

HUD_expanded$imonth3<-sprintf("%02d",HUD_expanded$imonth3)
HUD_expanded$iday3<-sprintf("%02d",HUD_expanded$iday3)

HUD_expanded<-HUD_expanded %>%
  mutate(iyear_quarter3=paste(iyear3,"-",imonth3,"-",iday3,sep = ""))

HUD_expanded$iDate3 <- as.Date(HUD_expanded$iyear_quarter3, "%Y-%m-%d")

HUD_expanded$iDate_correspond3 <- floor_date(HUD_expanded$iDate3,"quarter")

HUD_expanded<-HUD_expanded %>%
  mutate(w3_correspond=ifelse(rec_date==iDate_correspond3,1,0))

w3_correspond<-filter(HUD_expanded,w3_correspond=="1")
w3_correspond$delete<-duplicated(w3_correspond$aid)
w3_correspond<-filter(w3_correspond,delete==FALSE)

HUD_expanded<-HUD_expanded %>%
  mutate(non_resp3=if_else(is.na(iDate_correspond3) &
                           rec_date=="2001-04-01" |
                           rec_date=="2001-07-01" |
                           rec_date=="2001-10-01" |
                           rec_date=="2002-01-01" |
                           rec_date=="2002-04-01",1,0))

w3_nonresponse<-filter(HUD_expanded,non_resp3=="1")
w3_nonresponse$delete<-duplicated(w3_nonresponse$aid)
w3_nonresponse<-filter(w3_nonresponse,delete==FALSE)

###########
## Wave 4 ##
###########

HUD_expanded<-merge(HUD_expanded,w4_interview,by="aid",all.x = TRUE)

HUD_expanded$imonth4<-sprintf("%02d",HUD_expanded$imonth4)
HUD_expanded$iday4<-sprintf("%02d",HUD_expanded$iday4)
```

```
HUD_expanded<-HUD_expanded %>%
  mutate(iyear_quarter4=paste(iyear4,"-",imonth4,"-",iday4,sep = ""))

HUD_expanded$iDate4 <- as.Date(HUD_expanded$iyear_quarter4, "%Y-%m-%d")

HUD_expanded$iDate_correspond4 <- floor_date(HUD_expanded$iDate4,"quarter")

HUD_expanded<-HUD_expanded %>%
  mutate(w4_correspond=ifelse(rec_date==iDate_correspond4,1,0))

w4_correspond<-filter(HUD_expanded,w4_correspond=="1")
w4_correspond$delete<-duplicated(w4_correspond$aid)
w4_correspond<-filter(w4_correspond,delete==FALSE)

HUD_expanded<-HUD_expanded %>%
  mutate(non_resp4=if_else(is.na(iDate_correspond4) &
                           rec_date=="2007-04-01"|
                           rec_date=="2008-01-01" |
                           rec_date=="2008-04-01" |
                           rec_date=="2008-07-01" |
                           rec_date=="2008-10-01" |
                           rec_date=="2009-01-01",1,0))

w4_nonresponse<-filter(HUD_expanded,non_resp4=="1")
w4_nonresponse$delete<-duplicated(w4_nonresponse$aid)
w4_nonresponse<-filter(w4_nonresponse,delete==FALSE)
```

References

Bigback, Kristyn M., Megan Hoopes, Jenine Dankovchik, Elizabeth Knaster, Victoria Warren-Mears, Sujata Joshi, and Thomas Weiser. 2015. "Using Record Linkage to Improve Race Data Quality for American Indians and Alaska Natives in Two Pacific Northwest State Hospital Discharge Databases," *Health Services Research* 50 (Suppl 1): 1390–1402. doi: 10.1111/1475-6773.12331.

Blakely, Tony, and Claire Salmond. 2002. "Probabilistic Record Linkage and a Method To Calculate the Positive Predictive Value," *International Journal of Epidemiology* 31 (6): 1246–1252.

Center on Budget and Policy Priorities (CBPP). 2017. "Policy Basics: Public Housing." https://www.cbpp.org/research/policy-basics-public-housing.

Chantala, Kim, William D. Kalsbeek, and Eugenio Andraca. 2004. "Non-Response in Wave III of the Add Health Study." Add Health User Guide Series. https://www.researchgate.net/publication/268254454_Non-response_in_Wave_III_of_the_Add_Health_study.

Eggers, Frederick J. 2017. *Characteristics of HUD-Assisted Renters and Their Units in 2013*. Washington, DC: U.S. Department of Housing and Urban Development, Office of Policy Development and Research. https://www.huduser.gov/portal/sites/default/files/pdf/Characteristics-HUD-Assisted.pdf.

Garvey Wilson, Abigail L., Charles W. Hoge, Dennis McGurk, Jeffrey L. Thomas, Julie C. Clark, and Carl A. Castro. 2010. "Application of a New Method for Linking Anonymous Survey Data in a Population of Soldiers Returning from Iraq," *Annals of Epidemiology* 20: 931–938.

Gress, Taryn, Seungjong Cho, and Mark Joseph. 2017. *HOPE VI Data Compilation and Analysis*. Washington, DC: U.S. Department of Housing and Urban Development, Office of Policy Development and Research. https://www.huduser.gov/portal/sites/default/files/pdf/HOPE-VI-Data-Compilation-and-Analysis.pdf.

Hanlon, James. 2017. "The Origins of the Rental Assistance Demonstration Program and the End of Public Housing," *Housing Policy Debate* 27 (4): 611–639. https://doi.org/10.1080/10511482.2016.1262445.

Harris, Kathleen Mullan, Carolyn Tucker Halpern, Eric A. Whitsel, Jon M. Hussey, Ley A. Killeya-Jones, Joyce Tabor, and Sarah C. Dean. 2019. "Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health)," *International Journal of Epidemiology* 48 (5): 1415–1415k. https://doi.org/10.1093/ije/dyz115.

Kalsbeek, William D., Carolyn B. Morris, and Benjamin J. Vaughn. 2001. *Effects of Nonresponse on the Mean Squared Error of Estimates from a Longitudinal Study*. Presented at the 2001 Joint Statistical Meetings, Atlanta, GA, August 6–8. http://www.amstat.org/sections/SRMS/proceedings/y2001/Proceed/00194.pdf.

McClure, Kirk. 2018. "Length of Stay in Assisted Housing," *Cityscape* 20 (1): 11–38.

https://www.huduser.gov/portal/publications/length-of-stay.html.

National Center for Health Statistics (NCHS). 2016. *NCHS-HUD Linked Data Analytic Considerations and Guidelines*. Hyattsville, MD: National Center for Health Statistics: 51. https://www.cdc.gov/nchs/data/datalinkage/nchs_hud_analytic_considerations.pdf.

———. 2019. *NCHS-HUD Linked Data: Analytic Considerations and Guidelines*. Hyattsville, MD: National Center for Health Statistics: 28. https://www.cdc.gov/nchs/data/datalinkage/nchs_hud_analytic_considerations.pdf.

National Center for Health Statistics (NCHS), Office of Analysis and Epidemiology (OAE). 2019. *A Primer on HUD Programs and Associated Administrative Data*. Hyattsville, MD: National Center for Health Statistics, Office of Analysis and Epidemiology. https://www.cdc.gov/nchs/data-linkage/hud-methods.htm.

Pendall, Rolf, and Leah Hendey. 2013. *A Brief Look at the Early Implementation of Choice Neighborhoods*. Washington, DC: Urban Institute. https://www.urban.org/sites/default/files/publication/24126/412940-a-brief-look-at-the-early-implementation-of-choice-neighborhoods_1.pdf.

Thalji, Lisa, P. Perry, N. Grilley, B. Nichols, and Roger Tourangeau. 1997. *The Prospective Longitudinal Study of Adolescent Health In-School Component*. Final Report. Chicago, IL: National Opinion Research Center (NORC) and the University of Chicago.

Tromp, Miranda, Anita C. Ravelli, Gouke J. Bonsel, Arie Hasman, and Johannes B. Reitsma. 2011. "Results from Simulated Data Sets: Probabilistic Record Linkage Outperforms Deterministic Record Linkage," *Journal of Clinical Epidemiology* 64 (5): 565–572.

Udry, J. Richard, and Kim Chantala. 2003. "Missing School Dropouts in Surveys Does Not Bias Risk Estimates," *Social Science Research* 32 (2): 294–311.

U.S. Department of Housing and Urban Development (HUD). n.d.a. "About HOPE VI." https://www.hud.gov/program_offices/public_indian_housing/programs/ph/hope6/about.

———. n.d.b. "About IMS/PIC." https://www.hud.gov/program_offices/public_indian_housing/systems/pic/about.

———. n.d.c. "Moving to Work (MTW) Expansion." https://www.hud.gov/program_offices/public_indian_housing/programs/ph/mtw/expansion.

———. n.d.d. "Picture of Subsidized Households." https://www.huduser.gov/portal/datasets/assthsg.html.

———. n.d.e. "Voluntary Conversion." https://www.hud.gov/program_offices/public_indian_housing/centers/sac/vc.

U.S. Interagency Council on Homelessness (USICH). 2013. *PHA Guidebook to Ending Homelessness.* https://www.usich.gov/resources/uploads/asset_library/PHA_Guidebook_Final.pdf.

Webb, Michael D., Kirstin P. Frescoln, and William M. Rohe. 2016. "Innovation in US Public

Housing: A Critique of the Moving to Work Demonstration," *International Journal of Housing Policy* 16 (1): 111–124.

Winkler, William E. 2016. "Chapter 2: Probabilistic Linkage." In *Methodological Developments in Data Linkage,* edited by Katie Harron, Harvey Goldstein, and Chris Dibben. New York: John Wiley & Sons, Ltd.: 8–35.

Zhang, Yujia, Bruce Cohen, Maurizio Macaluso, Zi Zhang, Tonji Durant, and Angela Nannini. 2012. "Probabilistic Linkage of Assisted Reproductive Technology Information with Vital Records, Massachusetts 1997–2000," *Maternal and Child Health Journal* 16 (8): 1703–1708.

Zhu, Ying, Yutaka Matsuyama, Yasuo Ohashi, and Soko Setoguchi. 2015. "When To Conduct Probabilistic Linkage vs. Deterministic Linkage? A Simulation Study," *Journal of Biomedical Informatics* 56: 80–86.

U.S. Department of Housing and Urban Development
Office of Policy Development and Research
Washington, DC 20410-6000

May 2022