



# Evolving availability and standardization of patient attributes for matching

Yu Deng<sup>1</sup>, Lacey P. Gleason<sup>1</sup> , Adam Culbertson<sup>1</sup>, Xiaotian Chen<sup>2</sup> , Elmer V. Bernstam<sup>3,4</sup> ,  
Theresa Cullen<sup>5</sup> , Ramkiran Gouripeddi<sup>6</sup> , Christopher Harle<sup>7,8</sup> , David F. Hesse<sup>9</sup>,  
Jacob Kean<sup>10</sup> , John Lee<sup>11</sup> , Tanja Magoc<sup>12</sup> , Daniella Meeker<sup>13</sup> , Toan Ong<sup>14</sup>,  
Jyotishman Pathak<sup>15</sup>, Marc Rosenman<sup>16</sup> , Laura K. Rusie<sup>17</sup> , Akash J. Shah<sup>18</sup> , Lizheng Shi<sup>19</sup> ,  
Aaron Thomas<sup>20</sup> , William E. Trick<sup>21</sup> , Shaun Grannis<sup>8</sup>, Abel Kho<sup>1,\*</sup> 

<sup>1</sup>Center for Health Information Partnerships, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, United States

<sup>2</sup>Statistical Innovation Group, Data and Statistical Sciences, AbbVie, Inc, North Chicago, IL 60064, United States

<sup>3</sup>School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States

<sup>4</sup>Division of General Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States

<sup>5</sup>Pima County Health Department, Tucson, AZ 85714, United States

<sup>6</sup>Clinical and Translational Science Institute and Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84108, United States

<sup>7</sup>Department of Health Policy and Management, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN 46202, United States

<sup>8</sup>Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN 46202, United States

<sup>9</sup>Hesse Foot and Ankle Clinic, SC, Eau Claire, WI 54751, United States

<sup>10</sup>VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System and University of Utah, Salt Lake City, UT 84148, United States

<sup>11</sup>Edward Hospital, Naperville, IL 60540, United States

<sup>12</sup>Integrated Data Repository Research Services, Clinical and Translational Science Institute, University of Florida, Gainesville, FL 32609, United States

<sup>13</sup>Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT 06510, United States

<sup>14</sup>Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States

<sup>15</sup>Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, United States

<sup>16</sup>Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611, United States

<sup>17</sup>Howard Brown Health, Chicago, IL 60640, United States

<sup>18</sup>Nuvance Health, Danbury, CT 06810, United States

<sup>19</sup>Department of Health Policy and Management, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA 70112, United States

<sup>20</sup>North Carolina Translational and Clinical Sciences Institute, School of Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

<sup>21</sup>Center for Health Equity & Innovation, Cook County Health, Chicago, IL 60612, United States

\*Corresponding author: Center for Health Information Partnerships (CHIP), Northwestern University Feinberg School of Medicine, 625 N. Michigan Ave, Suite 1500, Chicago, IL 60611, United States. Email: [Abel.Kho@nm.org](mailto:Abel.Kho@nm.org)

## Abstract

Variation in availability, format, and standardization of patient attributes across health care organizations impacts patient-matching performance. We report on the changing nature of patient-matching features available from 2010–2020 across diverse care settings. We asked 38 health care provider organizations about their current patient attribute data-collection practices. All sites collected name, date of birth (DOB), address, and phone number. Name, DOB, current address, social security number (SSN), sex, and phone number were most commonly used for cross-provider patient matching. Electronic health record queries for a subset of 20 participating sites revealed that DOB, first name, last name, city, and postal codes were highly available (>90%) across health care organizations and time. SSN declined slightly in the last years of the study period. Birth sex, gender identity, language, country full name, country abbreviation, health insurance number, ethnicity, cell phone number, email address, and weight increased over 50% from 2010 to 2020. Understanding the wide variation in available patient attributes across care settings in the United States can guide selection and standardization efforts for improved patient matching in the United States.

**Key words:** patient matching; record linkage; electronic health records (EHRs); demographic attributes; data standardization; data completeness; data collection; interoperability.

## Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 encouraged the widespread adoption and meaningful use of electronic health records (EHRs).<sup>1</sup> Despite widespread use of EHR systems, interoperability between systems lags. According to the Office of the National Coordinator for Health Information Technology (ONC), interoperability is the “ability of two or more systems to exchange health information and use the information once it is received”.<sup>2</sup> The lack of interoperability between health care organizations has hampered care coordination, health information exchange, and efficient patient care and created fragmented silos of digital patient data.<sup>3</sup> The need for efficient data exchange for tracking and reporting during the COVID-19 pandemic underscored the importance of these challenges. A necessary component for interoperability and integration of patient records is optimization of patient matching.<sup>4</sup> Patient-matching methods typically rely on a combination of patient identity features from the EHR such as first name, last name, gender, social security number (SSN), and address.<sup>5</sup> Ideally, linkage variables should be unique, accurate, complete, and consistent across sites and time. However, in real-world practice, matching variables are often captured in varying formats in health care organizations and health information systems.<sup>6</sup> In addition, erroneous entry, data missingness, and information updates are common for many demographic variables. Understanding the availability of matching variables and the variability in their collection among health care organizations is a critical first step to improving patient matching.

While the US Core Data for Interoperability (USCDI) recommends specific data elements and standards to be used for interoperability purposes, it is unclear how well this is adopted in clinical practice.<sup>7</sup> Previous work<sup>8</sup> studied 36 patient attributes from 2005–2014 across 9 health care facilities and found that matching variables, such as first name, last name, date of birth (DOB), and gender/sex, were highly collected across locations and time. This study is a continuation and expansion of this prior work. In this study, we established an expert panel to identify patient attributes of interest. We investigated 63 data attributes (see [Table S1](#)), the standards used for their collection, and their availability from 2010 to 2020 across a diverse group of health care provider organizations in the United States.

## Data and methods

Our study aimed to measure how health care provider organizations collect patient demographic attributes that could be used for patient-matching algorithms. This project was submitted to the Northwestern University Institutional Review Board and determined not to be human subjects research. We first convened an expert panel to compile a list of patient demographic variables of interest and a list of proposed care settings to include in subsequent phases of the study. Second, we developed a questionnaire that included questions about which demographic variables were collected by each health care provider organization, how these variables were collected, and whether they were used for cross-provider patient matching. Third, we asked a subset of 20 organizations to complete a query to determine the availability of each of these patient demographic elements in their EHR systems from 2010 to 2020. For consistency, we developed

Structured Query Language (SQL) pseudocode and instructions on how to query aggregated data from SQL databases. These materials were sent to the 20 sites to ensure consistent implementation of the query.

## Expert panel

We convened a meeting of experts on patient demographics and record linkage in Washington, DC, on September 25, 2019. The morning sessions included presentations on the building blocks of interoperability; expert perspectives from industry, academia, and government; a facilitated discussion on policy implications; and a deep dive on US Postal Service (USPS) standards. The afternoon sessions included a facilitated discussion on putting theory into practice that involved brainstorming and identifying the crucial data elements that should be considered for patient matching as well as a discussion about proposed care settings and contacts for recruitment. The list of proposed sites for recruitment was based on care setting, populations served, rural vs urban location, and region of the country (see [Table S2](#)).

## Questionnaire

We designed and deployed a questionnaire using REDCap electronic data-capture tools hosted at Northwestern University.<sup>9</sup> Starting in May 2020, the project team directly contacted health care organizations, beginning with the list of proposed care settings and contacts developed during the expert panel discussions, and shared the questionnaire link. Although the initial intention was to use quota sampling to recruit participants matching the profiles of proposed care settings, due to the difficulty of recruitment during the initial stages of the COVID-19 pandemic, individual outreach was supplemented with other recruitment approaches. A project description with the questionnaire link was sent out through newsletters of professional organizations for health information management, rural health, and primary care; posted to the website for the Center for Health Information Partnerships (CHIP); and shared through CHIP’s Twitter account.

## Query development and analysis

To better understand the completeness of each attribute, each site reported the counts and percentages of patients with demographic attribute values by year between 2010 and 2020. Sites were asked to return (1) overall counts and percentage availability from 2010 to 2020 (inclusive) and (2) yearly counts and percentage availability. The overall count is defined as the total number of unique patients who had any visits, including virtual and in-person visits, from 2010 to 2020. The overall percentage availability is defined as the number of unique patients who have the patient attribute of interest available at any time from 2010 to 2020 divided by the overall count from 2010 to 2020. The yearly counts include the total number of unique patients whose last visits were in that individual year. For example, if patient A had visits in 2010, 2012, and 2014, patient A is only included in the yearly count for 2014. We used the year of the last visit to (1) avoid repeatedly counting the same patients with multiple visits and (2) maintain consistency between attributes that have historical records and those that only have the most recent record. The yearly percentage is defined as the number of unique patients with nonmissing values divided by its corresponding

yearly count. Eqn (1) also provides the definition of yearly percentage.

$$P_{i,t} = \frac{x_{i,t}}{N_t}, \tag{1}$$

where  $P_{i,t}$  is the yearly percentage for attribute  $i$  and year  $t$ .  $N_t$  is the total number of unique patients whose last visit was in year  $t$ .  $x_{i,t}$  is the number patients with nonmissing values for attribute  $i$  at any time up to year  $t$  among the  $N_t$  patients.

Only patients having at least 1 visit to the participating health care provider organization from 2010 to 2020 and who were between 18 and 89 years of age at time of extraction were included. Sites that did not have any data in a given year reported percentage availability of these data as 0%.

### Statistical analysis

For each patient attribute, minimum, maximum, and median percentage availability were calculated across the 20 sites. Median was chosen because it is more robust against outliers compared to mean. Trends in attribute availability across time were assessed using Cochran-Armitage tests. Statistical analyses were performed using Python 3.0.1 and R 4.0.0. Plots were generated using the “Seaborn” package in Python.<sup>10</sup> Cochran-Armitage tests were performed using the “DescTools” package in R.<sup>11</sup>

## Results

### Expert panel results

Sixteen individuals with expertise in patient matching attended the panel meeting. These attendees included experts in health information management and health informatics from government, academia, independent social research organizations, professional societies, and industry. The key output of the panel meeting was the creation of a list of demographic variables of interest and a list of proposed care

settings for project participants. The outputs of this meeting served as the foundation for the subsequent questionnaire development as well as identification of a set of geographically and demographically diverse care sites.

### Questionnaire results, current time period

We received 40 responses to our questionnaire from June 2020 through June 2022. Two responses were excluded from analysis because one was a duplicate response and one was an entirely blank response. As a result, 38 organizations were included in this analysis. Among the 38 organizations, there were 10 respondents (26%) from the Northeast, 7 respondents (18%) from the South, 10 respondents (26%) from the Midwest, and 11 respondents (29%) from the West. By care setting, there were 14 (37%) academic health centers, 8 (21%) integrated delivery systems, 5 (13%) community hospitals, 5 (13%) independent clinics, 1 (3%) long-term care facility, and 5 (13%) organizations that identified as “other”. Participants who identified their organization’s care setting as “other” included a tribal health clinic, a community center offering behavioral health services, a pediatric stand-alone hospital, a health center controlled network, and a group of health department clinics. By location, 32 respondents (84%) were from urbanized areas of 50 000 or more people, 1 respondent (3%) was from an urban cluster of at least 2500 and less than 50 000 people, and 5 respondents (13%) were from rural areas. However, participants were only given the ability to select 1 option for location, so this may not be representative of the true distribution for participating organizations with multiple sites, especially integrated delivery systems. The following proportions of questionnaire respondents indicated that patients identifying as Hispanic or Latino/a (74% of questionnaire respondents); Black or African American (68%); Asian (55%); lesbian, gay, bisexual, transgender, queer, intersex, and asexual (LGBTQIA) (53%);

**Table 1.** Count and percentage of questionnaire respondents that collected each demographic variable and used each demographic variable for cross-provider patient matching.

Variable	Collected variable	Used for matching	Not used for matching	Unknown if used for matching
Name	38 (100%)	26 (68.4%)	5 (13.2%)	7 (18.4%)
DOB	38 (100%)	27 (71.1%)	4 (10.5%)	7 (18.4%)
Address	38 (100%)	23 (60.5%)	8 (21.1%)	6 (15.8%)
Phone Number	38 (100%)	17 (44.7%)	13 (34.2%)	7 (18.4%)
Birth sex	36 (94.7%)	17 (44.7%)	11 (28.9%)	8 (21.1%)
Email	36 (94.7%)	4 (10.5%)	21 (55.3%)	11 (28.9%)
Ethnicity	35 (92.1%)	4 (10.5%)	20 (52.6%)	10 (26.3%)
Health insurance ID Number	34 (89.5%)	8 (21.1%)	15 (39.5%)	11 (28.9%)
Race	33 (86.8%)	6 (15.8%)	17 (44.7%)	9 (23.7%)
SSN	32 (84.2%)	16 (42.1%)	9 (23.7%)	7 (18.4%)
Language	30 (78.9%)	3 (7.9%)	20 (52.6%)	7 (18.4%)
Marital status	30 (78.9%)	0 (0%)	22 (57.9%)	8 (21.1%)
Gender identity	22 (57.9%)	4 (10.5%)	10 (26.3%)	7 (18.4%)
Occupation	18 (47.4%)	0 (0%)	15 (39.5%)	3 (7.9%)
Picture or image	18 (47.4%)	1 (2.6%)	12 (31.6%)	5 (13.2%)
Sexual orientation	15 (39.5%)	0 (0%)	11 (28.9%)	4 (10.5%)
Driver’s license number	11 (28.9%)	1 (2.6%)	6 (15.8%)	3 (7.9%)
Place of birth	9 (23.7%)	1 (2.6%)	5 (13.2%)	3 (7.9%)
Income	7 (18.4%)	0 (0%)	4 (10.5%)	3 (7.9%)
Tribal identity	3 (7.9%)	0 (0%)	2 (5.3%)	1 (2.6%)
Eye color	1 (2.6%)	0 (0%)	0 (0%)	1 (2.6%)
Other biometrics	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Data are presented as  $n$  (%). Source: Authors’ analysis of data from questionnaire. Abbreviations: DOB, data of birth; SSN, social security number.

immigrants (47%); American Indian or Alaskan Native (24%); Native Hawaiian or Other Pacific Islander (24%); and Middle Eastern or North African (18%) were well represented in their organization's patient population.

Table 1 shows the number and percentage of questionnaire respondents who collected each of the demographic variables as well as the number who used each demographic variable for cross-provider patient matching. All of the questionnaire respondents indicated that they collected name, DOB, address, and phone number. No respondents collected other biometrics, which included biometrics that include picture or image. Date of birth, name, current address, SSN, birth sex, and phone number were most commonly used for cross-provider patient matching. For many of the variables, some questionnaire respondents were unsure if that variable was used for cross-provider patient matching.

### Date of birth

All 38 respondents (100%) indicated that they collected the DOB. The most common format used was MM-DD-YYYY (month-day-year). When recording a month or a day with only a single digit, 28 respondents (74%) indicated that they put a leading zero. Twenty-seven respondents (71%) indicated that they used DOB for cross-provider patient matching, while 7 respondents (18%) indicated that it was unknown whether DOB was used for cross-provider patient matching. Formats used by respondents for collection of DOB are shown in Table S3.

### Name

All 38 respondents (100%) indicated that they collected names. The most commonly collected name types were first name/given name (92%), last name/family name/surname (89%), and middle name (including middle initial) (74%). The majority of respondents also collected preferred name (55%) and suffix (50%), while around one-third of respondents collected nickname (39%), previous last name(s) (34%), and previous first name(s) (32%). The types of names collected by questionnaire respondents and characters that can be included in names are shown in Figures S1 and S2.

### Address

All 38 respondents (100%) indicated that they collected address. The majority of respondents (55%) indicated that they used the USPS standard. Eight respondents (21%) indicated that they collected prior addresses. Twenty-three respondents (61%) indicated that they used current address for cross-provider patient matching, 4 respondents (11%) indicated that they used prior address for cross-provider patient matching, 8 respondents (21%) indicated that they used neither current nor prior addresses for cross-provider patient matching, and 6 respondents (16%) did not know if addresses were used for cross-provider patient matching. Standards used by respondents for collection of addresses are shown in Table S4 and address elements collected by respondents are shown in Figure S3.

### Social security number

Thirty-two respondents (84%) reported collecting SSN. In terms of the standard used for collection, 24 respondents (63%) indicated area number–group number–serial number, 4 respondents (11%) indicated serial number (last 4 digits), and 4 respondents (11%) indicated “other”. Among those who indicated “other”, 1 respondent indicated that they

collected SSN in a free-text format; 1 respondent indicated that they collected full SSN, last 4 digits, or a default number for those who don't know or refuse to provide SSN; 1 respondent indicated that they collected SSN to verify patient match if necessary; and the fourth respondent indicated that the format used was unknown. Of the 32 respondent organizations that collected SSN, 28 indicated that SSN is auto-formatted in their systems, meaning that the spaces or dashes are already there and do not need to be manually entered. Sixteen respondents (42%) indicated that SSN is used for cross-provider patient matching, while 7 respondents (18%) did not know if SSN was used for cross-provider patient matching.

### Birth sex

Thirty-six respondents (95%) indicated that they collected birth sex. Seventeen respondents (45%) indicated that birth sex is used for cross-provider patient matching, while 8 respondents (21%) did not know if birth sex was used for cross-provider patient matching. Table 2 shows the standards used for recording birth sex.

### Gender identity

Twenty-two respondents (58%) indicated that they collected gender identity. Four respondents (11%) indicated that gender identity was used for cross-provider patient matching, while 7 respondents (18%) did not know if gender identity was used for cross-provider patient matching. These respondents were asked what standard was used for recording gender identity. Table 3 shows these responses.

### Phone number

All 38 respondents indicated that they collected phone numbers. Twenty-seven respondents (71%) indicated that they

**Table 2.** Standards or answer choices used by questionnaire respondents for collection of birth sex.

Standard or answer choices used for collection of birth sex	Count
Male, Female, Unknown	14
HL7 Version 3 (V3) standard value sets for administrative gender	7
Blank	4
Male, Female	3
1 – Female 2 – Male 3 – Unknown 950 – Nonbinary 951 – X Another list for Sex assigned at birth are 1 – Female 2 – Male 3 – Unknown 4 – Not recorded on birth certificate 5 – Chose not to disclose 6 – Uncertain 999 – Other	1
Date is normalized to M (male), F (female), I (intersex), A (ambiguous), or U (unknown)	1
Female Male Nonbinary Other Unknown X	1
female, male, unknown, not recorded on birth certificate, choose not to disclose, uncertain	1
It is recorded as gender and some people change throughout their medical history. The gender may also change accordingly.	1
M, N, F, NULL	1
No standard	1
Patient's Sex assigned at birth Male Female Unknown Not recorded on birth certificate Choose not to disclose Uncertain	1
Sex Assigned At Birth: Male, Female, Intersex, Decline to Answer	1
Standard value sets	1

Source: Authors' analysis of data from questionnaire. Abbreviations: HL7, health level seven.

**Table 3.** Standards or answer choices used by questionnaire respondents for collection of gender identity.

Standard or answer choices used for collection of gender	Count
Blank	18
1 – Female 2- Male 3- Transgender Female/Male-to-Female 4 – Transgender Male/Female-to-Male 5 – Other 6 – Choose not to disclose	1
Choose not to disclose, Female, Male, Other, Something else, Transgender Female/Male-to-Female, Transgender Male/Female-To-Male	1
Female Male Transgender Female/Male to Female Transgender Male/Female to Male Nonbinary Other Chose not to disclose	1
Female, male, transgender female/male-to-female, transgender male/female-to-male, other, choose not to disclose, nonbinary	1
Female, Male, Transgender Female, Transgender Male, Other	1
Female, Male, Transgender Female/Male-to-Female, Transgender Male/Female-to-Male, Other, Choose not to disclose	1
Gender Identity: Male/Man, Female/Woman, Trans Male/Trans Man, Trans Female/Trans Woman, Genderqueer/ Gender nonconforming, Something Else, Decline to Answer	1
HL7	1
Is your gender identity different than your sex assigned at birth? Y/N Female Male Transgender Female/male to female Transgender Male/female to male Other Choose not to disclose Genderqueer/gender diverse Non-Binary Something else	1
M, N, F, NULL	1
Male Female Unknown	1
Male, female, both, neither, something else	1
Male, female, FTM, MTF, genderqueer, choose not to disclose, and additional category free text	1
Male, Female, Other, Transgender Female/Male to Female, Transgender Male/Female to Male, Choose not to disclose	1
Male/Female/Unknown/SOGI	1
Man Woman Transgender male/Trans man/Female-to-male Transgender female/Trans woman/Male-to-female Genderqueer/Non-binary Multiple gender categories Decline to answer Other	1
Patient's preference for self-identification (SOGI)	1
SNOMED	1
We ask what is on their birth certificate	1
WOMAN MAN GENDER NON-BINARY TRANSGENDER CISGENDER OTHER	1

Source: Authors' analysis of data from query. Abbreviations: HL7, health level seven; FTM, female-to-male; MTF, male-to-female; SNOMED, systematized medical nomenclature for medicine.

used the International Telecommunication Union (ITU)-T E.123 (02/2001) and ITU-T E. 164 standards, while 9 respondents indicated “other”. Of the 9 respondents who indicated “other”, 6 collected phone numbers in the following format ###-###-####, 1 indicated that it was a free-text field, 1 indicated that no standard was used, and 1 indicated that it was not consistent.

### Query results

Twenty questionnaire participant organizations also completed the query. These included 10 academic medical centers (50%), 1 community hospital (5%), 2 independent clinics (10%), 5 integrated delivery systems (25%), and 2 organizations from other care settings (10%). Participants were approximately evenly split by region, with 6 (30%) from the Midwest, 4 (20%) from the Northeast, 5 (25%) from the

South, and 5 (25%) from the West. The vast majority of query participant organizations (95%) were located in urbanized areas of 50 000 or more people. A total of 41 072 285 patients were seen at the 20 participating sites from 2010 to 2020. The median number of patients seen across sites from 2010 to 2020 was 1 246 510. Three sites had fewer than 100 000 patients, 4 sites had between 100 000 and 500 000 patients, 2 sites had between 500 000 and 1 million patients, 9 sites had between 1 million and 5 million patients, and 2 sites had greater than 5 million patients. [Table S5](#) and [Figure S4](#) show the distribution of the overall number of patients per site for organizations participating in the query. The median percentage of patients who identified as each of the following races alone across the 20 sites was 80.7% White, 11.9% Black or African American, 2.7% Asian, 0.4% American Indian or Alaska Native, and 0.2% Native Hawaiian or Other Pacific Islander. [Figure S5](#) shows the distribution of the known race alone populations across sites and demonstrates variation in the racial composition of the patient populations across participating sites.

### Overall availability

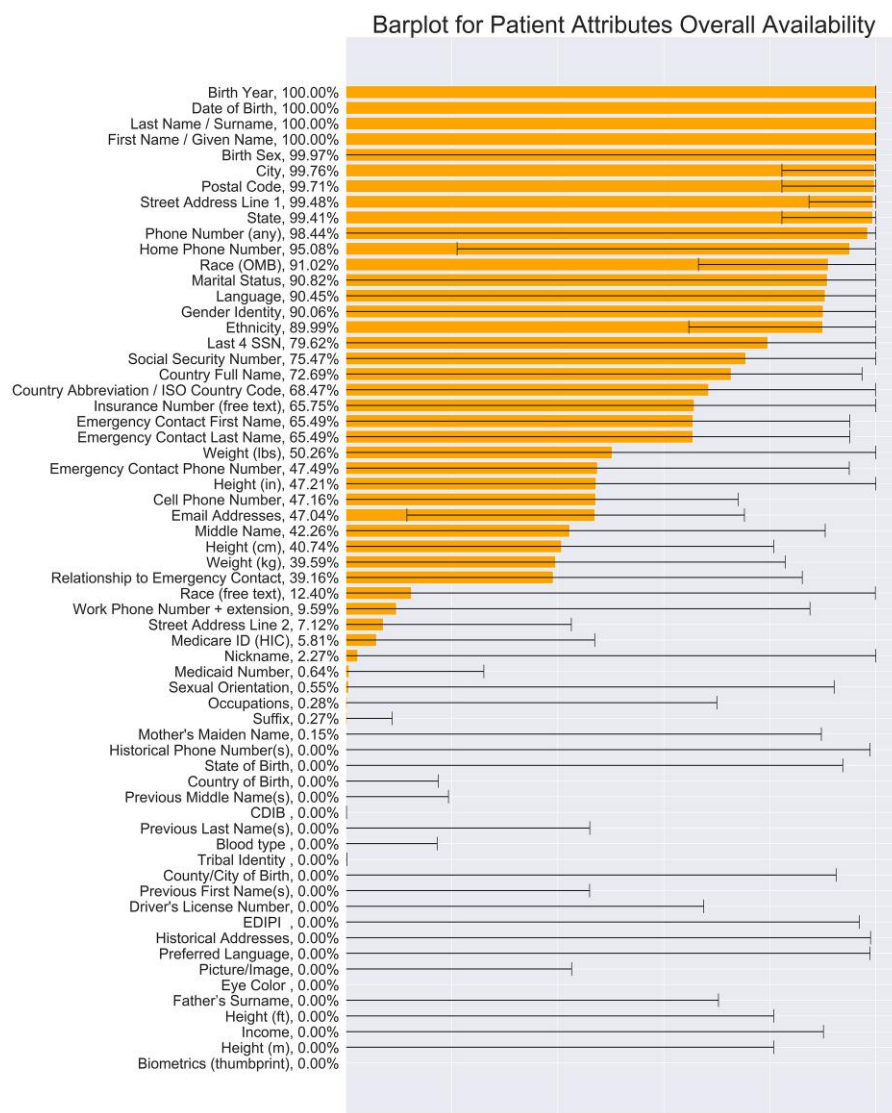
[Figure 1](#) shows the overall minimum, median, and maximum percentage availability for all 63 attributes across sites. Last name/surname, first name/given name, birth year, DOB, birth sex, city, postal code, state, street address line 1, phone number (any), and home phone number are highly available, with an overall median availability greater than or equal to 95%. Race (Office of Management and Budget [OMB]), marital status, language, gender identity, and ethnicity are often collected, with an overall median availability greater than or equal to 90%. The last 4 digits of the SSN (80%), country full name (73%), country abbreviation (68%) insurance number (66%), emergency contact first name (65%), emergency contact last name (65%), and weight (50%) are sometimes collected, with an overall median availability above or equal to 50%. Last, some patient attributes are rarely collected, with an overall availability percentage close to 1%. These patient attributes include sexual orientation, nickname/alias/preferred name, occupation, suffix, and tribal identity.

### Overall availability variation across sites

Patient attributes that have a median overall availability of 100% are also consistently collected across sites, except for birth sex (see [Figure S6](#)). Date of birth, birth year, last name/surname, and first name/given name were available for 100% of patients at all sites. However, birth sex availability varies greatly across sites. Interestingly, for patient attributes that are rarely collected (ie, overall availability ~0%), some of them are highly available in a small number of sites (see [Figure S7](#)). For example, 1 site had an overall availability percentage for sexual orientation above 90%. One site had an overall availability percentage of 100% for nickname/alias, and 1 site had an overall availability percentage of 100% for electronic data interchange personnel identifier (EDIPI), a unique personal identifier used by the Department of Defense. A similar pattern was observed for mother's maiden name, state of birth, income, and historical phone numbers.

### Attributes change across time

The Cochran-Armitage tests show that the availability percentage of all patient attributes, except for biometrics



**Figure 1.** Overall percentage of the availability of patient attributes. The bar plot shows the minimum, median, and maximum values of the overall availability of each patient attribute across sites. From bottom to top, the order of patient attributes is sorted by the value of median overall availability. The number to the right of each attribute name is the median overall percentage availability for that attribute. The left side of the black line on top of each bar represents the minimum value, and the right side of the black line on top of each bar represents maximum value among sites. Source: Authors' analysis of data from query. Abbreviations: CDIB, Certificate of Degree of Indian or Alaskan Blood; EDIPI, Electronic Data Interchange-Personal Identifier; HIC, Health Insurance Claim Number; ISO, International Organization for Standardization; OMB, Office of Management and Budget; SSN, social security number.

(thumb print) and eye color, changed between 2010 and 2020 ( $P < .00001$ ). Figure 2 shows the availability percentage of patient attributes from 2010 to 2020. Based on the trend of attribute availability percentage over the years, we categorized patient attributes into 4 groups: consistently high, consistently low, increasing, and decreasing.

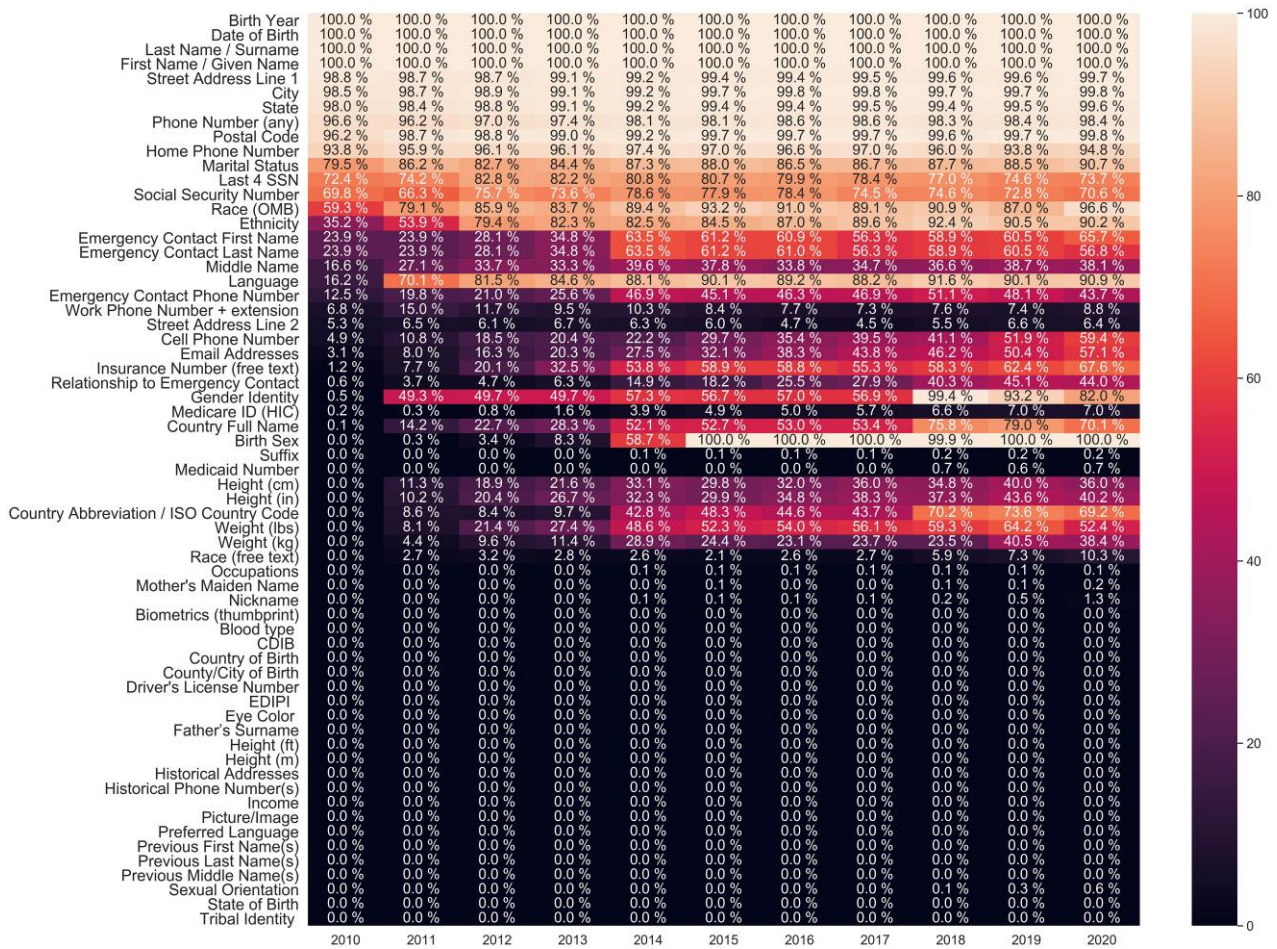
As shown in Figure 2, patient attributes that have consistently high availability across years are as follows: birth year, DOB, last name, first name, street address line 1, city, state, phone number (any), and postal code. The median availability of these attributes remains greater than or equal to 95% through all years.

Several patient attributes had consistently low availability over the years. These patient attributes include occupation, mother's maiden name, nickname, biometrics (thumb print), blood type, Certificate of Degree of Indian or Alaskan Blood (CDIB) number, country of birth, county/city of birth, driver's

license number, EDIPI, eye color, father's surname, height (feet), height (meters), historical addresses, historical phone number(s), income, picture/image, preferred language, previous first name(s), previous middle name(s), previous last name(s), state of birth, and tribal identity. All of these attributes had availability ranging from 0% to 1% over the 11 years studied.

The 10 patient attributes that have increased the most from 2010 to 2020 are as follows: birth sex, gender identity, language, country full name, country abbreviation, insurance number (free text), ethnicity, cell phone number, email address, weight (pounds), and relationship to emergency contact.

On the other hand, we observed that the availability of some patient attributes increased in the first couple of years and then decreased. These patient attributes include work phone number, SSN, and last 4 SSN digits. The availability of SSN and the last 4 digits of SSN increased from 2010 to 2014 but have



**Figure 2.** The median availability percentage of patient attributes from 2010 to 2020. The x-axis of the heat map represents time in years, and the y-axis represents each patient attribute. Each lattice color represents the availability percentage of that patient attribute in that specific year, with a darker color representing low availability and a lighter color representing high availability (see color bar). Source: Authors' analysis of data from query. Abbreviations: CDIB, Certificate of Degree of Indian or Alaskan Blood; EDIPI, Electronic Data Interchange-Personal Identifier; HIC, Health Insurance Claim Number; ISO, International Organization for Standardization; OMB, Office of Management and Budget.

declined since 2014. The availability of work phone number increased from 2010 to 2011 but has declined slightly from 2011 to 2019.

### Discussion

In this study, we analyzed the collection and use for patient matching of demographic attributes at health care organizations across the nation. On the questionnaire, the majority of respondents reported collecting conventional patient-matching attributes including name, DOB, address, and birth sex. We also saw that many respondents (95–100%) reported collecting contact information, including phone numbers and email addresses, and a significant number of respondents reported collecting identity-related attributes like birth sex, ethnicity, race, language, gender identity, and sexual orientation (~40–95%). Importantly, we found that there was significant variability in what standards are used for the collection of these demographic elements. Even for attributes that have relatively standard components, like name or DOB, we found that respondent organizations used a variety of different formats. This lack of standardization in attribute definitions and formatting likely hinders the successful use of these attributes for patient matching. Our results precede, but highlight,

the importance of recent national address standardization initiatives like Project US@,<sup>12</sup> recommended best practices for the collection of self-reported sexual orientation and gender identity (SOGI),<sup>13</sup> and recommendations for revising OMB's 1997 Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity.<sup>14</sup>

The query results demonstrate that some conventional patient-matching attributes, including DOB, birth year, first name/given name, and last name/surname, are highly available across all sites and across time. This is consistent with the results from a previous study.<sup>8</sup> Unlike other conventional patient-matching attributes, birth sex and gender identity are not consistently collected across sites. This may be due to some degree of conflation between these variables. In addition to conventional attributes that were highly available across the study period, we observed some emerging attributes, including birth sex, gender identity, language, country full name, country abbreviation, insurance number (free text), ethnicity, cell phone number, email address, weight (pounds), and relationship to emergency contact, all of which have increased greatly from 2010 to 2020. For some data attributes, such as sexual orientation and nickname, although the median overall availability percentages are comparatively low, a small number of sites had availability

percentages as high as 100%. This indicates that it is possible to consistently collect these data attributes, and positive outlier organizations may be well positioned to share best practices and promote alignment with federal government initiatives<sup>13</sup> for the successful collection of these emerging data elements.

### Strengths and limitations

There are some limitations to our study. First, we used non-probability sampling, which limits the generalizability and external validity of our findings. Compared with the list of proposed care settings developed during the expert panel discussions, we recruited fewer independent clinics and Federally Qualified Health Centers than planned. We found that these organizations, in particular, faced difficulties marshalling resources to participate in a project of this type in the face of resource constraints and competing priorities due to the coronavirus pandemic. The recruitment methods used likely bias the sample towards organizations with more resources to respond to the query and organizations that had a particular interest in demographic variable collection. We would expect that these organizations may have higher availability rates of demographic variables and may be more likely to use data-collection standards compared with the average health care provider organization in the United States. Thus, the observed results likely overestimate demographic data availability and adherence to standards.

Second, this study mainly focuses on data availability without evaluating data quality. This also likely results in overestimates of the availability of useful data as invalid data entry (eg, keystroke errors and inadvertent transposition of names and DOB) is common.<sup>15</sup> Because we did not collect the values of the attributes through our query due to data privacy and limited capability, we were not able to evaluate how the attribute values themselves change over time. In addition, co-occurrence may have an impact on the matching performance as well, depending on the matching algorithm used. All of the above issues are not the focus of this work but they might be interesting to investigate in the future.

Third, we have missing data in the first couple of years of the study period. Some sites used paper records for these years and could not query the completeness rate for variables in the paper records. Other sites transitioned between EHR systems during these years and were not able to query data from before the transition. We counted the availability percentage of patient attributes as zero for these years because these data could not be readily queried and used for patient matching.

Fourth, for the annual availability percentages, patients were included in the analysis only for the year of their last visit. Since different sites ran the query at different times, those who ran the query later may have had a higher proportion of their patients falling into the more recent years since there was more opportunity for the patients to have had a return visit. Due to differing capacities at these varied organizations, it was not feasible to have all sites run the query at the exact same time. Although the number of patients included in the analysis for each year is not balanced, we believe that only including patients whose last visit was in that year gives the most accurate picture of the data-collection practices of each organization at that point in time.

Last, the median of the percentages of White population alone (not including >1 race) among the 20 sites is 80.7%, which is slightly higher than US White-alone population

(75.8%) in the 2020 Census data. Even though the median percentage of the White population is higher than that in the Census data, we still observe that racial distribution varies across participating sites and several individual sites have more racially diverse patient populations.

These limitations reflect real-world challenges, such as competing priorities for organizational resources and lack of retention of historical attributes, that affect data availability and our ability to measure it across organizations over time. While it is important to consider these limitations, our study included health care organizations across the United States that were diverse in terms of geographic location, number of patients served, and racial distribution of patients and evaluated a comprehensive list of variables that could inform policy about patient attributes considered for incorporation into future matching algorithms.

### Policy implications—the shifting nature of identity and implications for matching

This study points to the shifting nature of identity as captured by patient attributes in health systems. These shifts are likely driven by societal changes in attitudes, behaviors, and policies as well as greater awareness of issues related to privacy and identity. While name and DOB remain consistently well captured, other attributes are not. For example, after an initial increase in capture, SSN has plateaued and even started to decrease over time, perhaps due to concerns around identity theft and sharing of this information. With a shift to mobile phone use<sup>16</sup> and increasing reliance on online business and social interactions, both cell phone number and email address are increasingly complete attributes that may be incorporated into future matching approaches. Our expert panel noted that, while various biometrics are promising future attributes for matching, their use and capture vary and are not widely implemented.

In both our questionnaire and data query analyses, we observed particularly widespread variation in the capture and standards used for gender and birth sex as attributes. Accordingly, we recommend particular attention being paid to which attribute is used for matching purposes, as differing values for sex or gender variables will likely result in poor match quality. The dissemination of current best practices for the collection of sex and gender identity will likely help with patient matching, given the inclusion of 1 of these constructs in most matching algorithms.<sup>15</sup> These standards will need to evolve as federal and non-federal agencies continue to conduct rigorous research and testing to ensure the accurate capture of and respect for the personal nature of these constructs.<sup>17</sup>

Beyond the general availability of attributes, we observed that, for other features commonly used for matching, seemingly minor variations in formatting (eg, in DOB or SSN) have the potential to significantly affect match rates. To maximize match rates, formatting standards should be agreed upon a priori, addressed by new or established matching approaches, or guided by policy.<sup>15</sup> Extant standards and policy, to be effective, also require adoption and awareness. Even for attributes with existing standards (eg, OMB race and ethnicity standards) we observed variation in real-world compliance and also in the format used for capture, with sites increasingly using free text (presumably to capture greater nuance). Questionnaire responses indicated confusion between birth



sex, gender identity, and sexual orientation as well as confusion between race and ethnicity. Some respondents provided answers that indicate they conflated these variables (eg, providing answer choices for ethnicity that reflect what is typically considered race). Encouragingly, however, capture rates increased significantly over the study period. Policies such as “Meaningful Use” of EHRs set targets for capture of these attributes and may have influenced completion rates.

Despite questionnaire respondents primarily serving in Health Information Management roles at health care organizations, several respondents were uncertain about standards used to record demographic variables and variables used for cross-provider patient matching. Additionally, some respondents indicated that they would need to confer with others across the organization to answer these questions, indicating that collaboration across individuals in various functions (eg, informatics, registration, clinical care) within an organization is necessary to sufficiently understand demographic variable collection.

Encouragingly, the USCDI version 3 already includes the following highly available and emerging attributes within its Patient Demographics/Information data class: First Name, Last Name, DOB, Race, Sex, Current Address, Phone Number, and Email Address. The ONC should continue to foster standardization of these attributes as they have done with address via Project US@. That same model and use of USCDI can extend to other attributes where standardization is lacking.

Another potential policy lever to improve standardization of demographic variables is for the Centers for Medicare and Medicaid Services to condition provider completion of Promoting Interoperability (PI) objectives on adherence to complete, standardized attribute matching. For example, 1 PI objective is health information exchange. If participating in direct matching or health information exchange, providers should use robust standards for attributes to meet the objective.

Perhaps 1 explanation for increased uptake is that payment policies that financially incentivize SOGI equity metrics have started to influence collection practices. State Medicaid programs are increasingly requiring collection of SOGI data for payment programs. Although these questionnaires are not yet well standardized,<sup>18</sup> payment policies are more likely than unenforced standards to influence data-collection practices and technology configurations, even for patients outside of the payment programs.

## Acknowledgments

This work was previously presented at the American Medical Informatics Association (AMIA) 2021 Annual Symposium, San Diego, November 2021, and the International Population Data Linkage Network (IPDLN) Conference, Edinburgh, Scotland, September 2022.

## Contribution statement

Y.D. and L.P.G. contributed equally.

## Supplementary material

Supplementary material is available at *Health Affairs Scholar* online.

## Funding

This work was supported by The Pew Charitable Trusts. This work was also supported in part by the National Center for Advancing Translational Sciences (NCATS) under award no. UL1TR000371.

## Conflicts of interest

Please see ICMJE form(s) for author conflicts of interest. These have been provided as supplementary materials.

## Notes

1. Henry J, Pylpchuk Y, Searcy T, Patel V. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015*. The Office of the National Coordinator for Health Information Technology; 2016.
2. Office of the National Coordinator for Health Information Technology. The path to interoperability 2013. Accessed March 21, 2023. [https://www.healthit.gov/sites/default/files/factsheets/onc\\_interoperabilityfactsheet.pdf](https://www.healthit.gov/sites/default/files/factsheets/onc_interoperabilityfactsheet.pdf)
3. Pew Charitable Trusts. Enhanced patient matching is critical to achieving full promise of digital health records; 2018. Accessed March 21, 2023. <https://www.pewtrusts.org/en/>.
4. Morris G, Farnum G, Afzal S, Robinson C, Greene J, Coughlin C. Patient identification and matching final report; 2014. Accessed March 21, 2023. [https://www.healthit.gov/sites/default/files/resources/patient\\_identification\\_matching\\_final\\_report.pdf](https://www.healthit.gov/sites/default/files/resources/patient_identification_matching_final_report.pdf)
5. Lee ML, Clymer R, Peters K. A naturalistic patient matching algorithm: derivation and validation. *Health Informatics J*. 2016; 22(4):1030-1044.
6. Grannis SJ, Xu H, Vest JR, et al. Evaluating the effect of data standardization and validation on patient matching accuracy. *J Am Med Inform Assoc*. 2019;26(5):447-456.
7. Office of the National Coordinator for Health Information Technology. United States Core Data for Interoperability (USCDI). HealthIT.gov2022; 2022. Accessed March 21, 2023. <https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi#draft-uscdi-v4>
8. Culbertson A, Goel S, Madden MB, et al. The building blocks of interoperability. A multisite analysis of patient demographic attributes available for matching. *Appl Clin Inform*. 2017;8(2): 322-336.
9. Harris PA, Taylor R, Minor BL, et al. The REDCap Consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208.
10. Waskom ML. Seaborn: statistical data visualization. *J Open Source Softw*. 2021;6(60):3021.
11. Signorell A, Aho K, Alfons A, Anderegg N, Aragon T, Arppe A, Baddeley A, Barton K, Bolker B, Borchers HW. DescTools: Tools for descriptive statistics. R package version 0.99; 2019;28:17.
12. Smiley C, Govan-Jenkins W. Project US@ Unified Specification for Address in Health Care 2013. Accessed March 21, 2023. <https://oncprojecttracking.healthit.gov/wiki/pages/viewpage.action?pageId=180486153>
13. Office of the Chief Statistician of the United States. Recommendations on the best practices for the collection of sexual orientation and gender identity data on federal statistical surveys 2023. Accessed March 21, 2023. <https://www.whitehouse.gov/wp-content/uploads/2023/01/SOGI-Best-Practices.pdf>
14. US Office of Management and Budget Interagency Technical Working Group on Race and Ethnicity Standards 2023. Accessed March 21, 2023. <https://spd15revision.gov/>
15. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *J Am Med Inform Assoc*. 2014;21(1):97-104.

16. Pew. Mobile fact sheet. 2021. Accessed March 21, 2023. <https://www.pewresearch.org/internet/fact-sheet/mobile/>
17. Federal Interagency Working Group. Evaluations of sexual orientation and gender identity survey measures: what have we learned? 2016. Accessed March 21, 2023. [https://dpcpsi.nih.gov/sites/default/files/Evaluations\\_of\\_SOGI\\_Questions\\_20160923\\_508.pdf](https://dpcpsi.nih.gov/sites/default/files/Evaluations_of_SOGI_Questions_20160923_508.pdf)
18. STATE HEALTH ACCESS DATA ASSISTANCE CENTER. Collection of Sexual Orientation and Gender Identity (SOGI) data: considerations for Medicaid and Spotlight on Oregon. 2021. Accessed March 21, 2023. <https://www.shvs.org/resource/collection-of-sexual-orientation-and-gender-identity-sogi-data-considerations-for-medicaid-and-spotlight-on-oregon/>