INTEGRATED DATA SYSTEM PERSON IDENTIFICATION: ACCURACY REQUIREMENTS AND METHODS[1]

Submitted to:

U.S. Department of Labor
Employment and Training Administration
Office of Policy Development and Research
Division of Strategic Planning and Performance
200 Constitution Avenue NW
Washington DC 20210

Submitted by:

Grace Fendlay
Director, Discretionary Grants
Division of Workforce Development and Adult Learning
Maryland Department of Labor, Licensing and Regulation
1100 North Eutaw Street
Baltimore, MD 21201

February 2012

[1] The authors of this report are Ting Zhang, PhD, Research Assistant Professor, and David Stevens, PhD, Research Professor, The Jacob France Institute, University of Baltimore. John Janak, Sang Truong and Jing Li participated in the data processing conducted for this research. The Institute is a sub-award recipient of Workforce Data Quality Initiative (WDQI) funds received by DLLR from the U.S. Department of Labor, Employment and Training Administration. The authors accept full and sole responsibility for the content of this report. Agreement or disagreement with the views expressed here should not be attributed to any other person or organization.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This report responds to a Workforce Data Quality Initiative (WDQI) challenge—the unreported quality of person identification (PI) features in many integrated data systems (IDS) that link confidential workforce, education and social services administrative records.

The importance of the PI topic reflects concern that many local k-12 education agencies do not collect student Social Security Numbers. Some conclude from this widespread omission that linkage of secondary student records with workforce data may be impossible. However, others have adopted *ad hoc* and commercial software solutions to bridge this gap. To date no standard record linkage method has been endorsed.

Will performance dashboards and research findings based on IDS information be accepted as trustworthy by individuals making important appropriation of funds, policy and program-level resource allocation decisions? Should IDS public-use releases be believed and acted upon?

A standard technical language is used in professional communication about PI topics. Record linkage can be pursued using *exact matching* or *statistical matching*. Within the *exact matching* portfolio are *deterministic* and *probabilistic* methods. And within the *deterministic* portfolio are *direct* and *hierarchical* methods.

A familiar first step among WDQI award teams is application of exact matching when two or more administrative data files each contains a SSN field. This first step is also the last step in some record linkage actions, which introduces selection bias threats, singly or in various combinations. Confirmation that a SSN has been issued, and is therefore valid, does not mean that the valid nine-digit SSN was issued to the person associated with this SSN in one or more administrative data files.

We completed a series of three record linkage steps: (1) determine what candidate identifiers are available in each administrative data set; (2) use Link Plus software to carry out multiple deterministic and probabilistic PI diagnostics; and (3) examine the potential matched pairs identified in step two, assigning each pair to one of three categories—match, non-match, or uncertain match.

Our intent has been to illustrate typical PI accuracy challenges that are found in administrative data files. These challenges occur over time within a single administrative data source and among different administrative data files.

Our diagnostic findings are not amenable to summary coverage. Sections 5 and 6 describe what steps we undertook and what we found.

Given our diagnostic findings to date: So what? If left unresolved, can a PI of unreported and perhaps unknown quality translate into unacceptable deficiencies in information, conclusions and recommendations that are released to stakeholders making important decisions about appropriation of funds, policies and program-level priorities?

PI accuracy is a necessary first step for successful integration of multiple administrative data sources. This is a universal requirement that applies to any and all attempts to link unit-record person specific administrative data sources.

Avoidance of stakeholder skepticism—rejection at worst—is within our collective control, but we need to take positive steps now to retain this control. Lost confidence is difficult to recover. We need to be out in front of this potential threat to realization of the return on past, current and future IDS investments.

We are not aware of an ongoing serious and sustained professional conversation about the criteria that are appropriate to define PI accuracy tolerances for specific applications. This conversation is needed because the community of practitioners does not know whether we are over- or under-investing in PI technologies and applications.

We encourage the U.S. Department of Labor, Employment and Training Administration WDQI leadership team to propose an appropriate forum—perhaps through the technical assistance resources of Social Policy Research Associates—to ensure immediate attention to the PI accuracy topic.

INTEGRATED DATA SYSTEM (IDS) PERSON IDENTIFICATION: ACCURACY REQUIREMENTS AND METHODS

I. INTRODUCTION

This report responds to a Workforce Data Quality Initiative challenge—the unreported quality of IDS person identification (PI)[2] features that link confidential workforce, education and social services administrative records.

The importance of the PI topic reflects concern that many local k-12 education agencies do not collect student Social Security Numbers. Some conclude from this widespread omission that linkage of secondary student records with workforce[3] data may be impossible. However, others have adopted *ad hoc* and commercial software solutions to bridge this gap. To date no standard record linkage method has been endorsed.

PI accuracy is one measure among many that determines the credibility that is accorded new IDS information. Other relevant, but not mutually exclusive, considerations include: the data sources themselves; administrative data field availability and quality; model specifications; statistical estimation methods; consistency of reported findings, conclusions and action recommendations with statistical results and complementary contextual information; and clarity of communication with targeted decision-making recipients.

Will performance dashboards and research findings based on P-20W integrated data system use be accepted as trustworthy by individuals making important appropriation of funds, policy and program-level resource allocation decisions?

Should IDS public use releases be believed and acted upon? The frontier of PI technologies and techniques continues to advance. Definition of a record linkage accuracy requirement threshold should be unique for each intended use. Medical and judicial case management actions, for example, require an extreme (high) accuracy level.

---

[2] Mulrow E. *et al*. (2011), *Final Report*: Assessment of the U.S. Census Bureau's Person Identification Validation System, Chicago, IL: NORC at the University of Chicago, 103 pp., is an accessible introduction to the topic; particularly Appendix A: Environmental Scan of Record Linkage Methods, pp. 51-94.

[3] There is no consensus or even clear convergence toward widespread shared agreement about the definition of *workforce* for IDS design and use.

What record linkage accuracy level is sufficient for statistical applications using a P-20W[4] state longitudinal data system? This question has not been answered yet, and we do not offer an answer here.

We advance understanding of the PI linkage quality of selected IDS component files by describing a series of diagnostic steps taken and resulting findings.

Our PI research continues. We look forward to communication with others to converge toward a standard PI practice or portfolio of practices that can be expected to improve future impacts on policy and program management decision-making.

2. ORGANIZATION OF THIS REPORT

Section 3 introduces standard PI terminology. Section 4 describes the administrative data files we used to carry out the pilot test research. Section 5 presents the PI steps we have completed to date. Section 6 reports our findings. Section 7 concludes with a description of how our progress on the PI topic intersects with and contributes to other ongoing WDQI projects.

3. PI TERMINOLOGY[5]

A standard technical language is used in professional communication about PI topics.[6] *Record linkage* is the basic action of connecting two or more sources of information that satisfies an explicit or implicit PI accuracy threshold.

Mulrow *et al.* (2011) includes a chart[7] from a 2010 Statistics Canada international methodology symposium workshop[8] that classifies available record linkage method choices. Record linkage can be pursued using *exact* matching or *statistical* matching. Within the *exact* matching portfolio are *deterministic* and *probabilistic* methods. And within the *deterministic* portfolio are *direct* and *hierarchical* methods.

---

[4] P-20W is the commonly accepted acronym for the time continuum from early childhood or pre-school (P) through postsecondary education (20) and/or participation in the workforce (W).

[5] Abbreviations are used in the remainder of this report for Person Identification (PI), Social Security Number (SSN), Workforce Data Quality Initiative (WDQI), State Longitudinal Data System (P-20W SLDS), and Integrated Data System (IDS).

[6] Mulrow E. *et al.* (2011), pp. 51-94, is an excellent recent example that we have drawn upon for this section of our report.

[7] Mulrow *et al.* (2011), *op cit*, p. 51.

[8] Fox, Karla and Stratychuk, Lori. (October 2010), *Proceedings of the Statistics Canada International Methodology Symposium*, Workshop 1: Record Linkage Methods; abstract available at http://www.statcan.gc.ca/conferences/symposium2010/work-atel-eng.htm#1. Abstracts of other relevant sessions, also covering record linkage topics, are available at http://www.statcan.gc.ca/conferences/symposium2010/abs-res-eng.htm#a34.

3.1 Exact matching

A familiar first step among WDQI award teams is application of exact matching when two or more administrative data files each contains a SSN field. This first step is also the last step in some record linkage actions, which introduces selection bias threats, singly or in various combinations:

- The SSN data field can include nine-digit sequences that do not satisfy known Social Security Administration issuance rules. Routine edit checks using these rules are practical and have been adopted by most professionals that work with SSNs for record linkage purposes.

- The SSN data field can contain less than nine-digits. It is unlikely that the same less than nine-digit string would appear in two administrative data sources associated with the same person, but this is possible. This is a first example of the need for caution and appropriate diagnostics, so an accurate match is not discounted too quickly.[9]

- A single accurate nine-digit SSN—that is, one that is known to have been issued by the Social Security Administration—may have been used by more than one person; and the mix of multiple users may differ over time within a single data file and among administrative data files that are of interest.

- Transposition of digits within a valid nine-digit SSN can result in hasty discard of what could be used to complete a successful exact match.

Deterministic matching

There are multiple deterministic matching methods. Criteria for selection of a preferred method from available options will typically be specific to a proposed application. How, and how much, does assured match accuracy matter? Matching complexity can be costly, so the method chosen should be proportional to the consequences of false positive and false negative results from matching.

Direct matching. Mulrow *et al.* (2011) describes *direct* matching as an all-or-nothing method that requires exact agreement on all identifier fields used. This match-merge approach can rely on a single identifier, such as a SSN, or multiple fields, such as full name, date of birth, gender, and location. Challenges encountered in adopting this method include missing information, transpositions, and erroneous entry of invalid information.

[9] Of course, if the partial nine-digit string appears in administrative data sources other than a UI wage record some desired uses of an integrated data system may still be impossible, depending upon the availability of other PI data fields.

Hierarchical matching. This exact match method offers positive strengths in the context of the challenge described at the outset of this report—omission of SSNs from an increasing percentage of k-12 student records. Hierarchical matching involves a sequence of steps that use different candidate identifier fields in each step. The suggested approach is to begin with the identifier field or combination of fields that are thought to be most reliable, making a match-no match judgment and then moving on to other steps to further populate the two classifications using the pool of remaining no matches from the previous step. Prior data field cleansing is urged to protect against final discard of records that should have been defined as matches, and would have been if an appropriate level of investment in data cleansing had been made.

Probabilistic matching

Unlike direct and hierarchical deterministic matching, the portfolio of probabilistic matching algorithms has a common feature—a defined statistical model is adopted to calculate an exact match probability.

Mulrow *et al.* (2011) describes this method:

> The advantage of probabilistic record linkage is that it uses all available identifiers to establish a match … and does not require identifiers to match exactly. Identifiers that do not match exactly are assigned a "distance" measure to express the degree of difference between files. Each identifier is assigned a weight and the total weighted comparison yields a score, which is used to classify records as linked, not linked, or uncertainly linked according to whether the probability of a match exceeds a certain threshold. (p. 58)

3.2 Statistical matching

Unlike the exact match methods described above, statistical matching does not use one or more unique identifier fields to accomplish linkage. Instead, pairings across data sets are based on similarity of chosen data fields. The quality of resulting matches can depend upon one's confidence in the specification of the matching rules.

The statistical matching method appears to have limited relevance for most WDQI IDS applications, but future comparisons of results obtained from multiple administrative data files with known unique entity identities may be worthwhile.

3.3 Additional data matching resources

In addition to the Mulrow *et al.* (2011) environmental scan of record linkage methods, which we found to be accessible for non-expert reading, there are many other treatments of the topic by highly respected professionals. Available abstracts of the 2010 Statistics Canada International Methodology Symposium sessions, described in footnote 7 above, are one example.

Herzog, T.N., Scheuren, F. and Winkler, W.E. (2007), *Data Quality and Record Linkage Techniques*, New York, NY: Springer reflects decades of collaborative work by the authors affiliated with the U.S. Census Bureau. And, for the serious professional, http://www.hcp.med.harvard.edu/statistics/survey-soft/docs/WinklerReclinkRef.pdf is a seven-page list of record linkage references prepared by W.E. Winkler.

4. DATA SOURCES USED

Four confidential administrative data file extracts were used to complete our pilot study diagnostics[10]:

- School District A[11] Data (SDAD) student record extracts, 1998—2010 (SDAD).
- Graduates Cohort X[12] (GCX) student record extracts, 2009; a smaller extract than SDAD.
- Maryland Department of Labor, Licensing and Regulation (DLLR) Job Training Partnership Act (JTPA) statewide participant record extracts, 1984-2000[13]; 181,000 unique records.
- Maryland Workforce Exchange (WFE) statewide participant record extracts that include Workforce Investment Act Title I and Job Service coverage, 2005-2009; 295,000 unique records.

Figure 1 illustrates the data sources used and the coverage timeline of each data source.

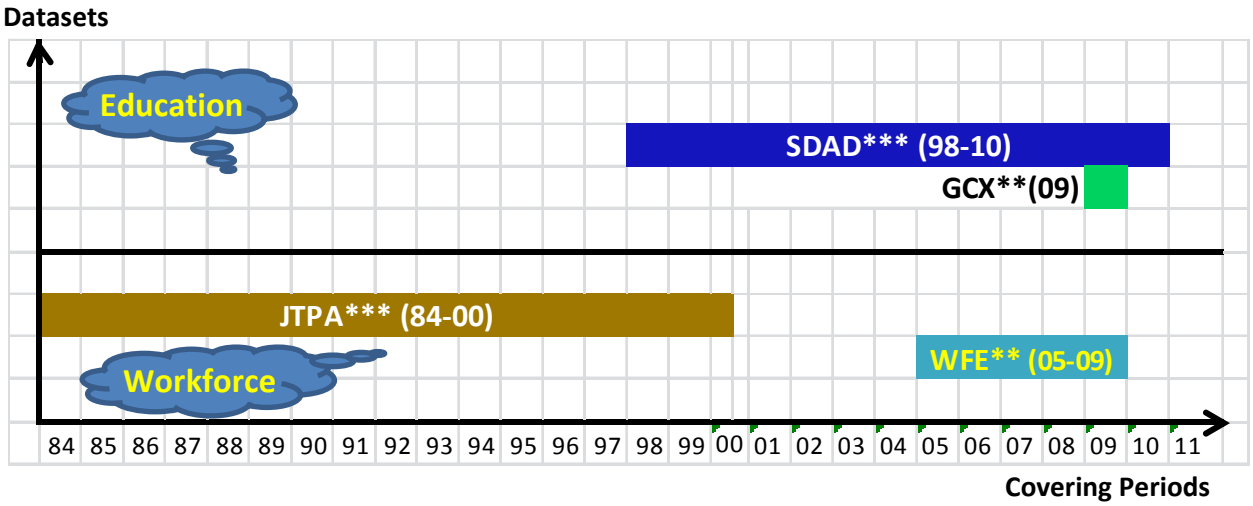
[10] The Jacob France Institute is affiliated with the University of Baltimore, which is a campus of the University System of Maryland. The Institute serves as an agent of multiple State agencies and local secondary and postsecondary education entities. In this technical support, research and evaluation capacity, authorized JFI staff members receive, maintain, process and use linked confidential administrative records for approved purposes. All confidential administrative data are received and processed in a secure environment. The data file extracts utilized to date for the described diagnostics are a subset of a larger number of confidential administrative data files maintained by or accessible to authorized JFI staff members for approved uses.

[11] We use "School District A" to avoid disclosure.

[12] Again, the specific dataset name is not disclosed.

[13] We found individuals identified as JTPA participants into calendar year 2004. These individuals were registered participants at the time of transition from JTPA to WIA administrative records in July 2000, so they remained in the administrative data base until inactivity or a recorded exit triggered the end of the reference spell of participation.

FIGURE 1: PROFILE OF DATA SOURCES AND TIME COVERAGE



**	SSN + Dob, or SSN + Name, or +gender (or race or education.)
***	SSN + Name + Dob + gender (or race, education or address).

5. PI DIAGNOSTIC STEPS

We combine deterministic matching and probabilistic matching to link the administrative data files defined in the previous section. Our goal is to identify PI quality issues related to the use of SSNs and other identifiers; singly in the case of SSNs, and in multiple combinations otherwise.

5.1 SSN validity

Our first step was to test SSN validity based on the Social Security Administration's monthly SSN issuance release. We repeat relevant points made in Section 3.1. Confirmation that a SSN has been issued, and is therefore valid, does not mean that the valid nine-digit SSN was issued to the person associated with this SSN in one or more administrative data files.

5.2 Matching methods

We completed a series of three record linkage steps:

1. Determine what candidate identifiers are available in each administrative data set.

2. Use Link Plus software to carry out multiple deterministic and probabilistic PI diagnostics.

3. Examine the potential matched pairs identified in Step 2, and assign each pair to one of three categories—match, non-match, or uncertain match.

5.2.1 Candidate identifiers

Candidate PI data fields found in two or more of the four data sources included in our pilot testing phase are:

- SSN
- Date of birth (DOB)
- Name (last, first, middle; full or partial)
- Gender
- Race and/or ethnicity
- Educational attainment
- Registration date

5.2.2 PI diagnostics completed using Link Plus software

Link Plus is a probabilistic record linkage software product originally designed to be used by cancer registries. However, Link Plus can be used with any type of data and has been used extensively across diverse research disciplines.

Link Plus enables a preliminary match between a 9 digit SSN in one file and a 4 digit SSN in a second file. If the last 4 digits of the 9 digit number are the same as the 4 digit number, the comparison pair will receive a higher preliminary score than when the last 4 digits are not the same in the two comparison files. This should be treated as a first step, not as conclusive evidence of a true match.

Link Plus also identifies specialized name and date matches, exact matches and enables phonetic matching on blocking variables[14] such as person name components or any variables for which pronunciation versus spelling helps to identify an individual record.

[14] For files with millions of records, the total of all possible comparison pairs is too large for practical computation. Blocking Variables are variables common to the two files that are used to 'block' (or partition) the two files. Only within these blocks are matching variables compared between the records. Blocking is a way to reduce the computing cost by portioning files into mutually exclusive and exhaustive blocks and performing comparisons only on records within each block.

Our first pilot diagnostic step, using one of the four administrative data files at a time, was to screen for individuals having the same recorded SSN but a different recorded DOB. For these cases we then added other available PI data fields to determine whether multiple data field matches justified tentative acceptance as a 'probable' match. This requires a decision-rule that determines whether the tentative acceptance should be followed by a costly manual review and assignment decision as either a 'probable' match or a non-match.

Our second pilot diagnostic step was to match across paired combinations of the four data files, using SSN or DOB as a blocking variable and the remaining PI fields as matching variables. The resulting pairs were saved for further manual review.

Giving highest initial priority to SSN accuracy, when available, we adopted a zero mismatch threshold as the criterion for deciding whether further diagnostics—statistical and/or manual—were required. This conservative exact match criterion is unlikely to be an appropriate choice as the final arbiter of acceptance or discard of a pair of records; but it is helpful to real-time case management decision-making.

5.2.3 Assignment of Step 2 pairs to matches, non-matches, or uncertain matches

The conservative second step procedure described in the previous subsection results in third step complexities and costs. Step 2 processing produced the following combinations of PI data fields in attempted matching across sequential pairings of two of the four administrative data files:

- Same SSN and similar surname, but different DOB.
- Same DOB and surname, but different SSN.
- Same DOB and surname, but missing SSN.
- Same DOB and similar surname, but different or missing SSN.
- Same DOB and surname, but different middle initial and SSN.
- Same DOB, surname and education, but different SSN.
- Same DOB, surname, race and gender, but different middle initial and SSN.
- Same DOB, registration date and education, but different SSN.
- Same DOB and registration date, but different education and SSN.
- Same DOB and education, but different registration date and SSN.
- Same DOB, race *or* gender, but different SSN.

Not all of the above candidate PI data field combinations were available in each of the four administrative data files used for pilot testing. Therefore, the actual combination of matching fields used in each processing step depended upon the particular draw from only two, or all four, data files.

Our basic decision rule for assignment as a match, probable match, or non-match was: When a potential matched pair has the same SSN and same DOB, we assign the pair as an exact match. However, when the potentially matched pair shares the same SSN, but different DOB, our assignment decision process became more complicated.

If a potentially matched pair also shares the same first, middle and surname, the probability that this pair identifies the same person is relatively high, particularly when the DOB is similar. However, if the potentially matched pair reveals different DOB and different name fields—even when the pair shares the same gender, race or education information— the probability that the pair identifies the same person is lower. For those pairs that share DOB and name, but different SSN, the likelihood that this is an acceptable match is not as high as those records with the same SSN, but higher than those pairs with the same DOB, different SSN and the same gender, race, or education information.

The remainder of this report profiles some of the complexities of administrative record quality for PI matching, and resulting impacts on the validity and reliability of public-use information that is drawn from components of a P-20W integrated data system.

6. FINDINGS

6.1 Introduction

Both workforce and education administrative data files were used in our pilot diagnostics. Our report of diagnostic findings begins with examples of PI matching results within a defined administrative data file over time. Abbreviations for each of the data files are used throughout the remainder of this section—JTPA, WFE, SDAD and GCX. The first two data files are DLLR workforce administrative sources, the third is a single public school district source, and the fourth includes a cohort of graduates.

6.2 Screening for duplicate records

We did not identify any records with the same SSN and no other shared identifiers. We did identify potentially matched pairs with the same DOB and name and/or same race or education information. SSN appears to be an effective unique identifier within each data file, but DOB combined with one or more other shared identifiers offers unclear evidence of potential duplication.

Our SDAD diagnostics found 44 true, or exact, matches; that is, paired records having the same SSN and DOB. Almost 170,000 pairs were identified with the same DOB, but different SSN and other available PI fields. Most of the pairs with only same DOB are likely to be different individuals. However, 35 of these pairs have the same DOB and same first and surname, but different middle initial and different SSN. An additional 7 pairs have the same DOB, surname and education, but different SSN.

In SDAD, it is not uncommon to see a 9-digit SSN starting with a "9" or a 7-digit SSN (shown in Tables 1 and 2). In this case, the string of digits is likely to be a different type of identification number.[15] If the matched pair includes other common fields, such as same DOB and surname, the chance that the pair identifies the same person increases. However, probabilistic matching alone does not guarantee a higher probability that the two potentially matched records pertain to the same individual.

Table 1 SDAD Data: Same DOB and Name, but Different SSN and MI

	SSN	DOB	First Name	Last Name	Middle Initials	College Graduate
Record 1	XXXXXXX	D (same)[16]	D (same)	D (same)		
Record 2	XXXXXXXXX	D (same)	D (same)	D (same)	A	

Table 2 SDAD Data: Same DOB, Name and Education, but Different SSN

	SSN	DOB	First Name	Last Name	Middle Initials	College Graduate
Record 1	3XXXXXX	D (same)	D (same)	D (same)		Yes
Record 2	2XXXXXXXX	D (same)	D (same)	D (same)		Yes

The JTPA diagnostic found 668 pairs identified with the same DOB, surname, race and gender, but different SSN. A manual scan found many of these pairs to have similar, but not identical, SSNs, as shown in Table 3. There is a relatively high probability that these pairs identify unique individuals.

The WFE diagnostic found over 2,000 pairs with the same DOB and education, or gender, or race, or registration date. These pairings did not give us a precise criterion for deciding whether this type of pairing identifies a unique individual.

[15] An Individual Taxpayer Identification Number (ITIN) is a tax processing number issued by the Internal Revenue Service. It is a nine-digit number that always begins with the number 9 and has a range of 70-88 in the fourth and fifth digit. Effective April 12, 2011, the range was extended to include 90-92 and 94-99 in the fourth and fifth digit, example 9XX-90-XXXX. IRS issues ITINs to individuals who are required to have a U.S. taxpayer identification number but who do not have, and are not eligible to obtain a Social Security Number (SSN) from the Social Security Administration (SSA). ITINs are issued regardless of immigration status because both resident and nonresident aliens may have a U.S. filing or reporting requirement under the Internal Revenue Code.
http://www.irs.gov/individuals/article/0,,id=222209,00.htm.l. An Alien Registration Number (A#) is also assigned to some individuals by the USCIS.

[16] D (same) means the field is the same. For disclosure avoidance reasons we do not report this field.

Table 3 JTPA Data: Same DOB, Name, Race and Gender, but Similar SSN

	SSN	DOB	First Name	Last Name	ethnicity	Gender
Record 1	XXXXXX0XX	D (same)	D (same)	D (same)	D (same)	D (same)
Record 2	XXXXXX9XX	D (same)	D (same)	D (same)	D (same)	D (same)

When SSN is missing or inconsistent, using a combination of other data fields sometimes helps to identify a person, but not always. If the other data fields, including DOB, first name, surname, middle initial, race/ethnicity, education and gender are all the same, but SSN in one potentially matched record is missing or is a different type of identifier, the probability that such a potentially matched pair is a true match could be high.

If a potentially matched pair receives a much higher matching score than a defined cutoff threshold value, shared common identifiers can help to replace a missing or erroneous SSN in one of the paired records. But, as we noted earlier, no standard cutoff threshold value has been defined for P-20W SLDS use. Also earlier in this report we expressed our opinion that the threshold value should be use-specific. Case management applications require a higher standard of PI accuracy than statistical applications. PI errors in transactions involving a person can result in unintended harm.

6.3 Matching education and 'workforce' data

Our diagnostics using the SDAD and WFE data files resulted in 37% of more than 12,000 potential matched pairs being true matches. The JTPA and SDAD diagnostic found 36% of 6,603 potential matched pairs to be true matches. The WFE and GCX diagnostic found only 18% of 809 potentially matched records to be true matches.

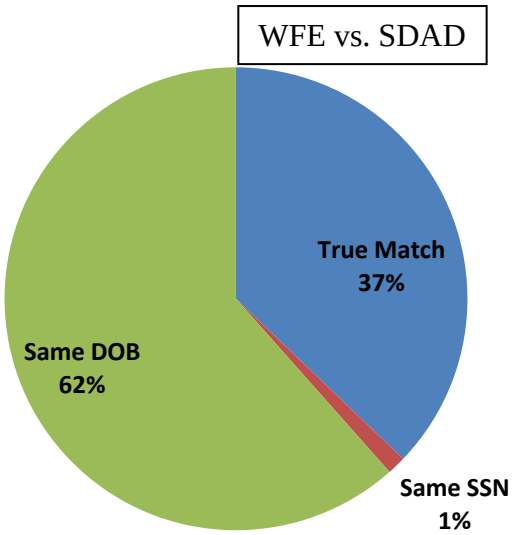
It is important to remember the time coverage and defined population for each of the four administrative data files we used for these pilot diagnostics (*Figure 1*). This reminder is particularly important with reference to a companion WDQI research project and forthcoming report that relies on PI accuracy to study individual participation in one or more public programs over time.

Older JTPA is a relatively small data file compared to the more recent and comprehensive WFE file that includes participants in more than one program—Workforce Investment Act Title 1 and Job Service in particular, but PI diagnostics covering both the JTPA and WFE files produced a similar percentage of true matches; the *numbers* of true matches differ.

For this pilot study, only 2009 GCX records were included. Therefore, GCX is a smaller data file than the single district SDAD file, which includes multiple school years.

It was no surprise to find only one potential matched pair between the pre-2001 JTPA and 2009 GCX files. Many more potential matched pairs were found in the GCX and WFE diagnostic. Figure 2 through Figure 4 show details of these diagnostics.

Figure 2 Matching Results Using WFE and SDAD Data Files



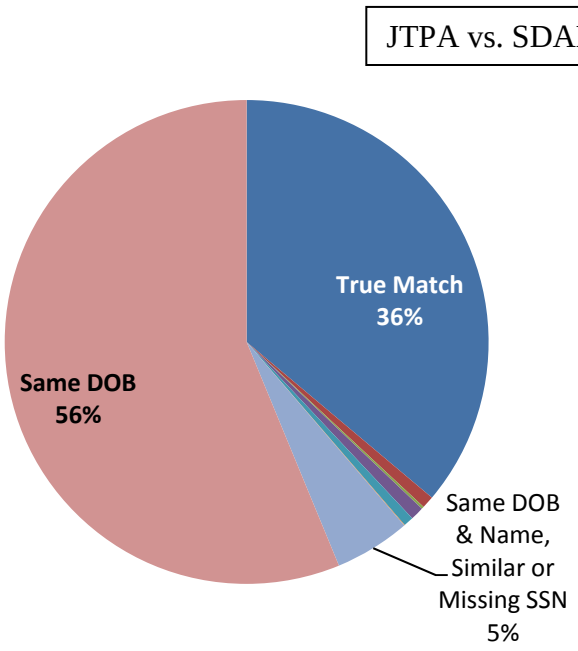
Potential matches	#
True Match	4,590
Same SSN Only	160
Same DOB Only	7,603
Total	12,353

One type of potential matched pair in our diagnostics completed to date identify records with the same SSN, but not necessarily common additional identifiers. If a potential matched pair shares an SSN and also an identical or very similar name, such as between JTPA and SDAD (*Figure 3*), the pair probably refers to one person.

Further diagnostic steps are required if the same SSN is associated with different DOB and different names (or race/ethnicity and gender). The same SSN may have been used by more than one person; a data entry error may have occurred; Or a name change may have happened. Resolution may require multiple additional diagnostic steps in such cases.

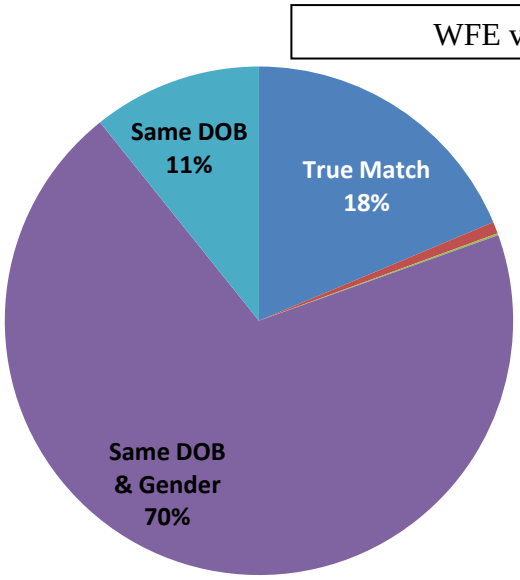
Figures 1 through 3 indicate that it is not unusual for a potential matched pair to share the same SSN but have different other fields. Linkage of multiple administrative data files allows cross checking of information to diagnose the roots of data quality concerns, in some cases leading to correction and improved assignment of match, no match, and probable match judgments.

Figure 3 Matching Results between JTPA Data and SDAD Data



potential matches	#
True Match	2,385
Same SSN Same Name	54
Same SSN Similar Name	11
Same SSN Only	59
Same DOB Similar or Missing SSN Similar Name	44
Same DOB Same Name Different SSN	3
Same DOB & Name Similar or Missing SSN	334
Same DOB Only	3,713
Total	6,603

Figure 4 Matching Results between WFE and GCX Data



potential matches	#
True Match	151
Same SSN & gender	6
Same SSN	1
Same DOB & Gender	564
Same DOB	87
Total	809

Using different combinations of SSN, names and DOB to match JTPA and SDAD offers additional insights. The combination of same DOB, same name but similar or missing SSN accounts for 5% of all matches. For instance, the name John Smith[17]is registered in both JTPA and SDAD with the same DOB, but the associated SSN is missing in the SDAD record.

A decision rule is needed to decide whether to accept this pairing as a sufficiently probable match to permit entry of the JTPA SSN into the SDAD record. Table 4 illustrates this case. This is a particularly interesting situation because JTPA time coverage is prior to the SDAD coverage. It is certainly possible that these two record sources refer to one person, but it is also possible that this is a case of a father and son without awareness of the Senior-Junior relationship. It is important to think expansively when PI accuracy matters.

Table 4 Pairs of Same DOB and Name, but Missing SSN across JTPA and SDAD Data

	SSN	DOB	First Name	Last Name
SDAD Record		D (same)	D (same)	D (same)
JTPA Record	XXXXXXXXX	D (same)	D (same)	D (same)

Other situations that we describe as being 'close to' true matches from JTPA-SDAD diagnostics are: (a) same SSN and same name; b) same SSN similar name; and c) same DOB, similar or missing SSN, and similar name. Table 5 summarizes these three situations.

For a case in situation a, the exact name shows up in both data sets with the same SSN. However, the recorded DOB is December 9, 1986 in the SDAD file, but December 29, 1986 in the JTPA file.

For situation b, the same SSN is shared between the potential matched pair, but one is associated with the name Cory Smith[18], while the other first name spelling is Smith. The DOB also differs by a single digit.

For situation c, two SSNs differ by one digit. With the same DOB, similar surname, and similar SSN, the probability is high that the pair identifies one person.

[17] Not an actual name found in the administrative data file.
[18] We use "Smith" to mask the actual last name to avoid disclosure.

Table 5 Summary of Potential Matches between JTPA and SDAD Data

Situation	SSN	DOB	Name
a	D (same)	1986120**9** vs. 1986122**9**	D (same)
b	D (same)	XXXX082**4** vs. XXXX082**3**	Cory Smith vs. Cor**e**y Smith
c	Different in One digit	D (same)	John[19] Mccormic vs. John Mccormic**k**

Inconsistency in each of these examples might be attributed to human reporting and/or data entry error, different ways to write a name, and a data field width that is not compatible with the same defined field in another data file; thus resulting in truncation errors. Truncation is a typical case where simple probabilistic matching alone does not generate satisfactory threshold to determine whether there is a true exact match.

Diagnostics conducted using WFE and GCX data files produced few interesting results. However some findings are worth noting. In Table 6, a case is listed with the same DOB and same gender, but the SSN differs by one digit. Using probabilistic matching software, this case may be accepted as a probable match, depending upon the specification of the cutoff point for acceptance, while there is a high possibility that this pair does not refer to the same individual.

Table 6 Case with Same DOB and Same Gender between WFE and GCX Data

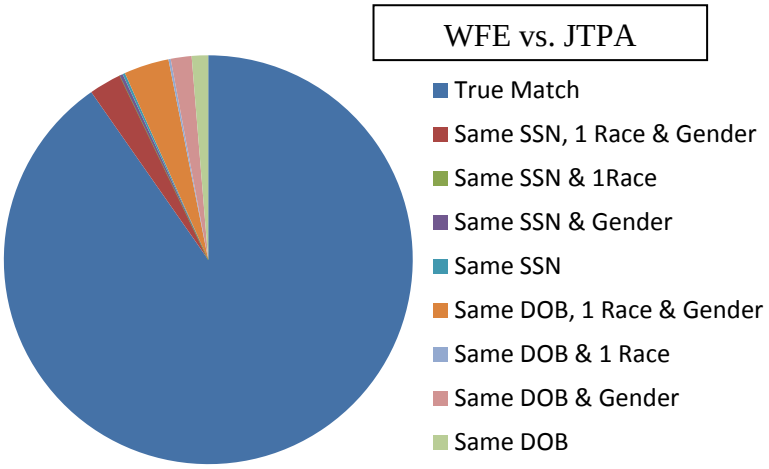
	SSN	DOB	Gender
Work Force Record	XXXXXXXX**2**	D (same)	D (same)
GCX Record	XXXXXXXX**4**	D (same)	D (same)

6.4 Matching education-education and workforce-workforce data files

Matching diagnostics conducted using two education data files, or two workforce data files, produces higher true match rate results than in the education-workforce pairings described in the previous subsection. The JTPA-WFE diagnostic produced 29,366 total matches, of which more than 90% are true matches. The two education data sets returned a predominant 99% true matches among all potential matched pairs. Figures 5 and 6 illustrate details.

[19] We use “John” to mask the actual first name to avoid disclosure

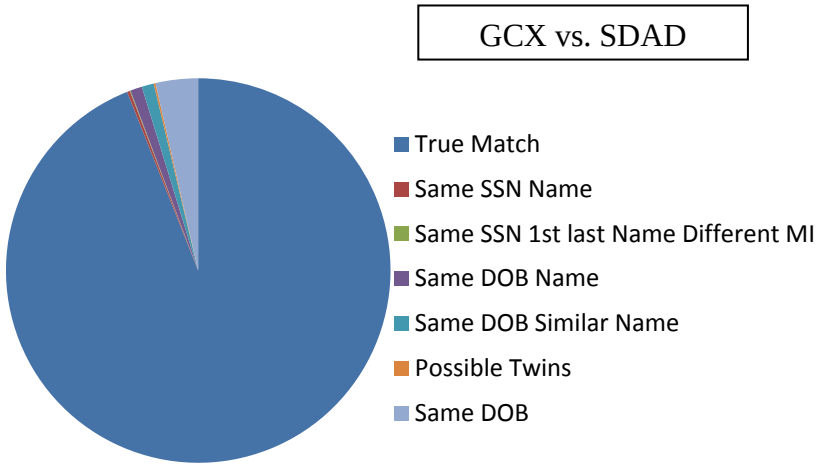
Figure 5 Matching Results between WFE and JTPA Data



potential matches	#
True Match	26,516
Same SSN , 1race & Gender	755
Same SSN & 1race	1
Same SSN & Gender	73
Same SSN Only	55
Same DOB, 1 Race & Gender	1,047
Same DOB & 1 Race	51
Same DOB & Gender	484
Same DOB Only	384
Total	29,366

Again, potential matched records with the same SSN but different other PI fields were found. This alerts us to be careful in hasty acceptance of SSN matches alone as sufficient evidence of a true, or exact, match.

Figure 6 Matching Results between SDAD Data and GCX Data



Potential matches	#
True Match	6832
Same SSN Name	17
Same SSN 1st last Name but Different MI	4
Same DOB Name	73
Same DOB Similar Name	73
Possible Twins	11
Same DOB	260
Total	6849

In the case of possible twins (shown in Table 7), we found that the potential matched pair has the same DOB, surname and middle initial, but different first name with the same initial. The SSN is different by one digit, but it is the final digit that is

strong evidence that two SSNs were issued at the same time and place. This is another case where limitations of probabilistic matching require extra caution.

***Table 7 Case of Possible Twins found between GCX and SDAD* Data**

	SSN	DOB	First Name	Last Name	Middle Initial
GCX Record	XXXXXXXX**5**	D (same)	KETRONE	D (same)	D (same)
SDAD Record	XXXXXXXX**6**	D (same)	KEVON	D (same)	D (same)

Table 8 illustrates another interesting case. The surname is different, but only because of truncation in the SDAD file. The SSNs are different, but this is because an incomplete SSN appears in the SDAD file. Still this pair is very likely to identify the same person.

Table 8 Case of Same DOB and Similar Name, but Different or Missing SSN between GCX and SDAD Data

	SSN	DOB	First Name	Last Name
GCX Record	220XXXXXX	D (same)	D (same)	**KALMANOVICH**
SDAD Record	220	D (same)	D (same)	**KALMANOV**

6.5 Summary of completed PI diagnostic findings

Our intent in this section has been to illustrate typical PI accuracy challenges that are found in administrative data files. These challenges occur over time within a single administrative data source and among different administrative data files.

An unanswered question up to this point is: So what? If left unresolved, can PI assignments of unknown quality translate into unacceptable deficiencies in information, conclusions and recommendations that are released to stakeholders making important decisions about appropriation of funds, policies and program-level priorities?

7. CONCLUDING REMARKS

7.1 Conclusions

PI accuracy is a necessary first step for successful integration of multiple administrative data sources. This is a universal requirement that applies to any and all attempts to link unit-record person specific administrative data sources. . A case-management diagnosis presented in this report illustrates the complexity of integration challenges and directions for further efforts.

There are multiple reasons why the thirteen-state WDQI has elevated aggressive pursuit of the PI topic to a new high priority:

1. Successful integration of 'workforce' administrative records with k-12 student records is often not possible relying on a common Social Security Number identifier alone.

2. Activation of the new Family Educational Rights and Privacy (FERPA) *Final Regulations*[20] in January 2012 will have an immediate impact on the number and types of requests for 'workforce'-education administrative data linkage.

3. Maturation of state P-20W longitudinal data systems has reached, or soon will reach, a tipping point after which stakeholder expectations about performance accountability reporting capabilities will accelerate.

4. The anticipated rising tide of performance accountability releases will trigger "How did you arrive at these reported findings?" questions, motivated by neutral curiosity in some cases, but by skepticism about accuracy in other cases.

5. A pervasive inability, or unwillingness, to answer queries about PI processing will pose a threat to overall acceptance of performance accountability measures and trends, and complementary research findings, based on IDS information.

Avoidance of stakeholder skepticism—rejection at worst—is within our collective control, but we need to take positive steps now to retain this control. Lost confidence is difficult to recover. We need to be out in front of this potential threat to realization of the return on past, current and future IDS investments.

[20] http://www.gpo.gov/fdsys/pkg/FR-2011-12-02/pdf/2011-30683.pdf.

The PI situation today can be simplified as follows:

- The portfolio of commercial software solutions to the PI challenge is expanding.

- These 'solutions' are often aggressively marketed in procurement contexts that constrain informed consideration of options and selection of an optimal product that is appropriate for the unique application intended.

- We are not aware of an ongoing serious and sustained professional conversation about the criteria that are appropriate to define PI accuracy tolerances for specific applications.[21] This conversation is needed because the P-20W community of practitioners does not know whether we are over- or under-investing in PI technologies and applications.

7.2 Next steps

We encourage the U.S. Department of Labor, Employment and Training Administration WDQI leadership team to propose an appropriate forum—perhaps through the technical assistance resources of Social Policy Research Associates—to ensure immediate attention to the PI accuracy topic.

We, and hopefully, others will continue to study whether and how decision-rules about PI accuracy impact our reported performance accountability measures and research findings. Having done so, we need to communicate our findings in forums and language that will be understood by non-experts. This is how we can improve our persuasiveness among those that make important appropriation of funds, policy and program-level management decisions.

[21] The fluid dynamics of the topic are such that this conversation may already be underway. The rapid changes occurring in medical technologies and service delivery logistics, for example, are motivating diverse commercial responses to PI challenges. The same can be said about innovation incentives in other sectors, including financial services and cyber security. A pertinent question is: Are ongoing conversations in other sectors relevant to the P-20W accuracy requirements; if so, how?