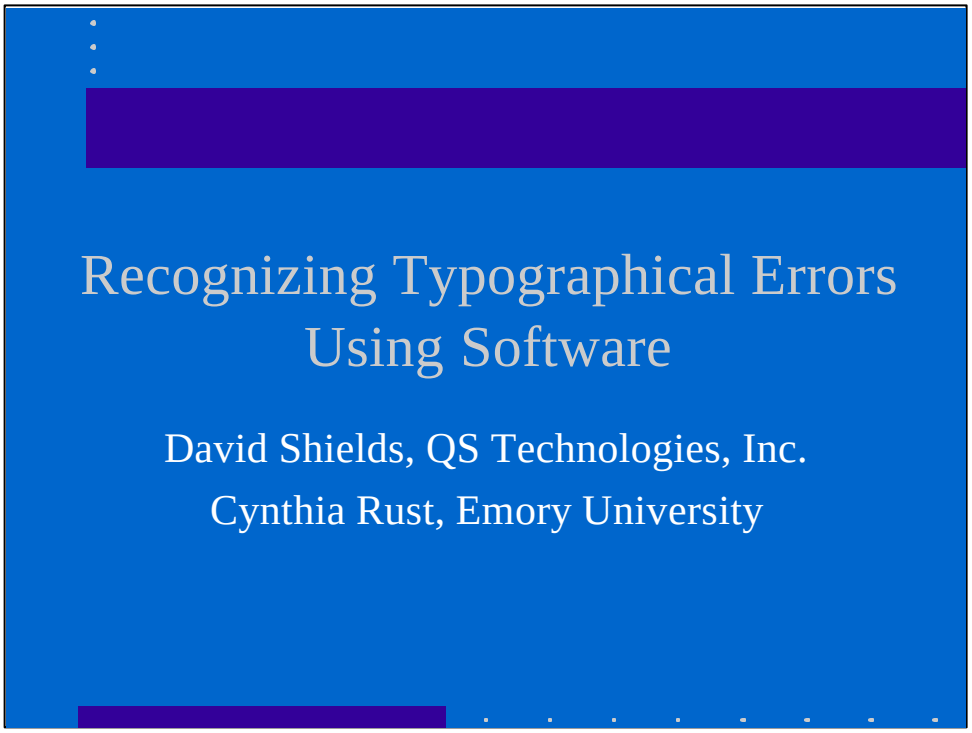
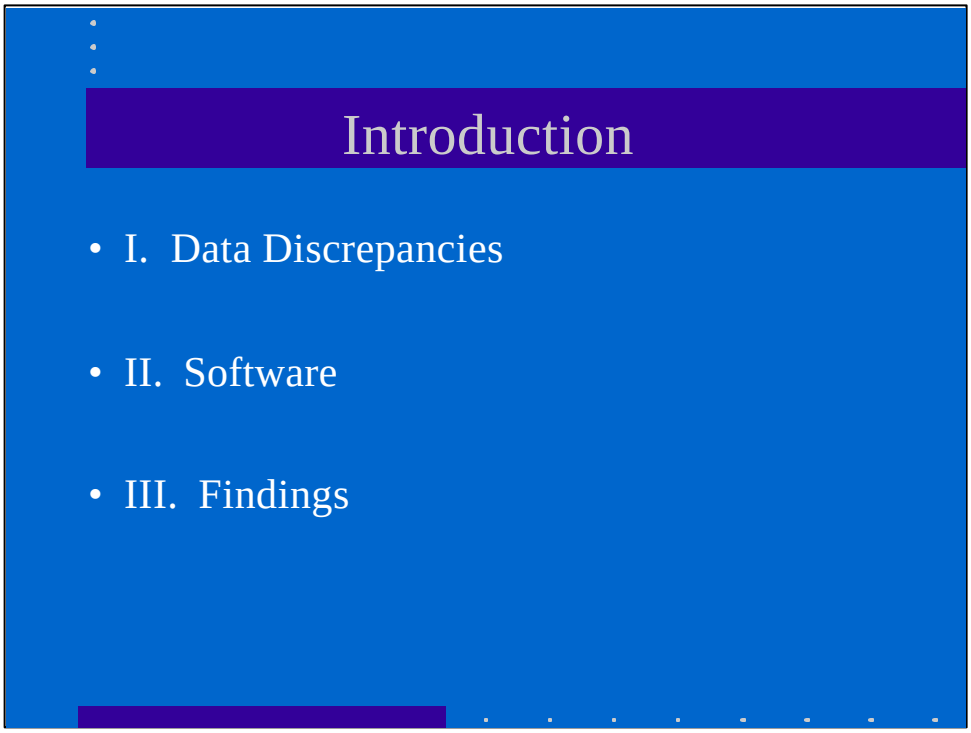


Most of the research presented in these slides was originally done by Emory University under contract to CDC, as part of an Immunization Registry De-duplication study.



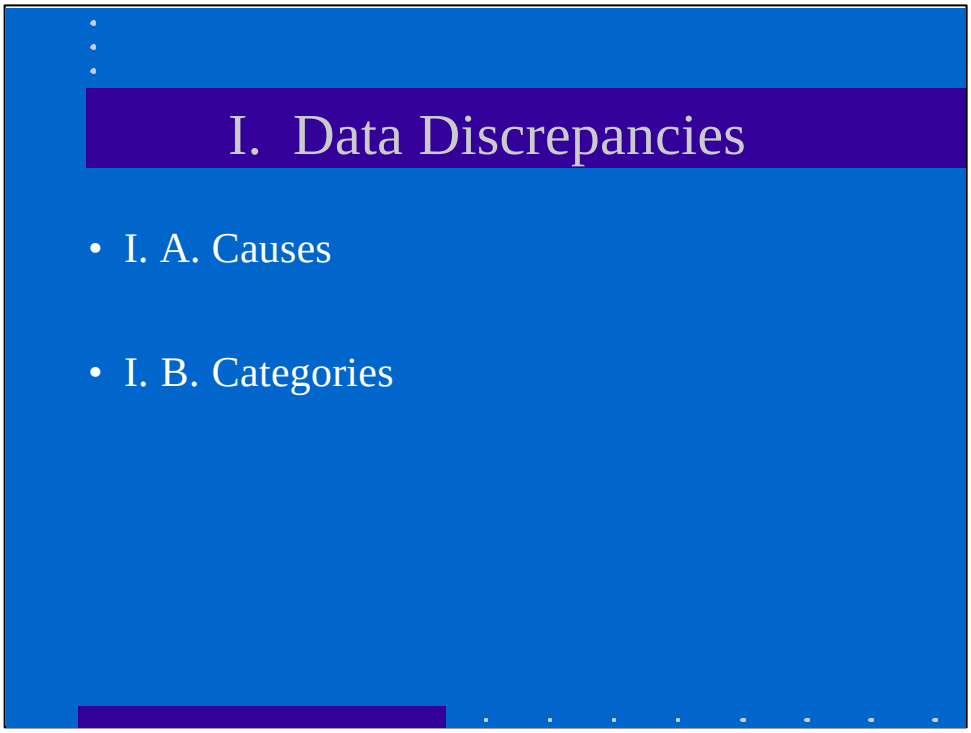
This presentation deals with discrepancies between data values when one person is represented by more than one record in an Immunization Registry.

Outline of this presentation is divided into 3 topics.

I. Classifications of Data Discrepancies (which will also be called **Typos**, Errors, or Discrepancies in the rest of this presentation).

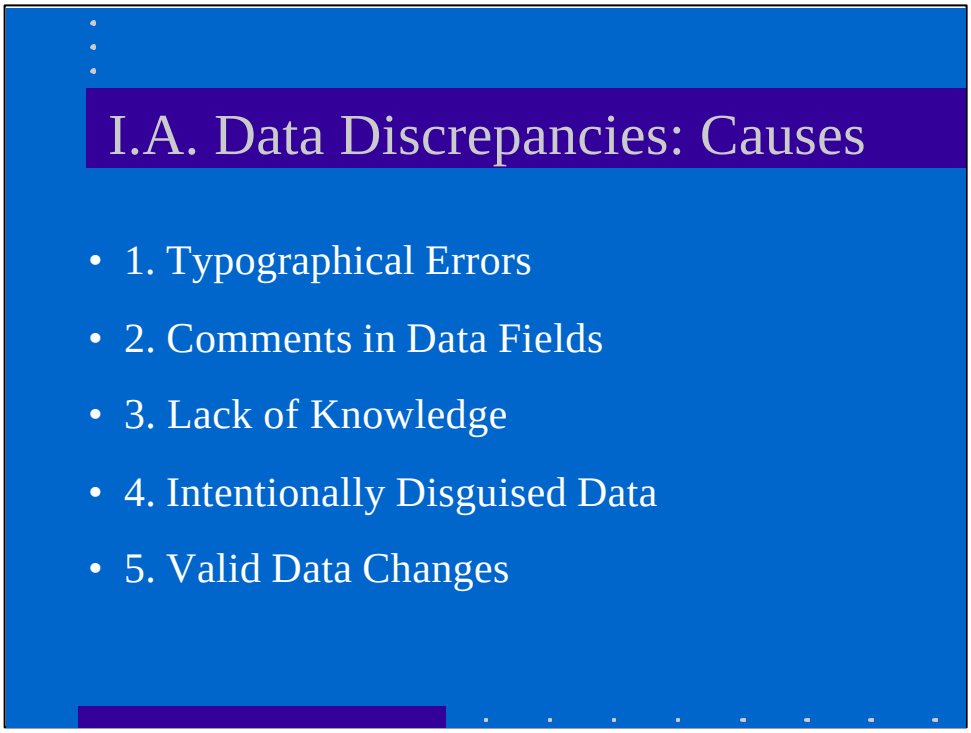
II. Structuring **software** to identify Data Discrepancies.

III. The **Findings** of the frequency of a variety of data discrepancies and typographical errors in the MATCH Immunization Registry.



I. A. We are better able to understand and classify discrepancies in data if we understand the **causes** of the discrepancies.

I. B. It is also useful for writing software to classify the discrepancies in such a way that we can produce algorithms to recognize different **categories** of data discrepancies and typographical errors.



1. Typographical Errors

In the narrowest definition, this refers to unintentional mistakes made during the transcription of data from one place to another.

For the purposes of this presentation, we will also talk about four other causes of data discrepancies, which are sometimes also lumped in with “typos”:

2. Comments in Data Fields

It is common for data entry clerks to work around limitations in their software by putting comments and messages in data fields. This would include things like “Baby Boy” as a first name, or “NFA” (No Forwarding Address) as a street address, or “Do Not Use, Duplicate Record” in a parent name field.

3. Lack of Knowledge

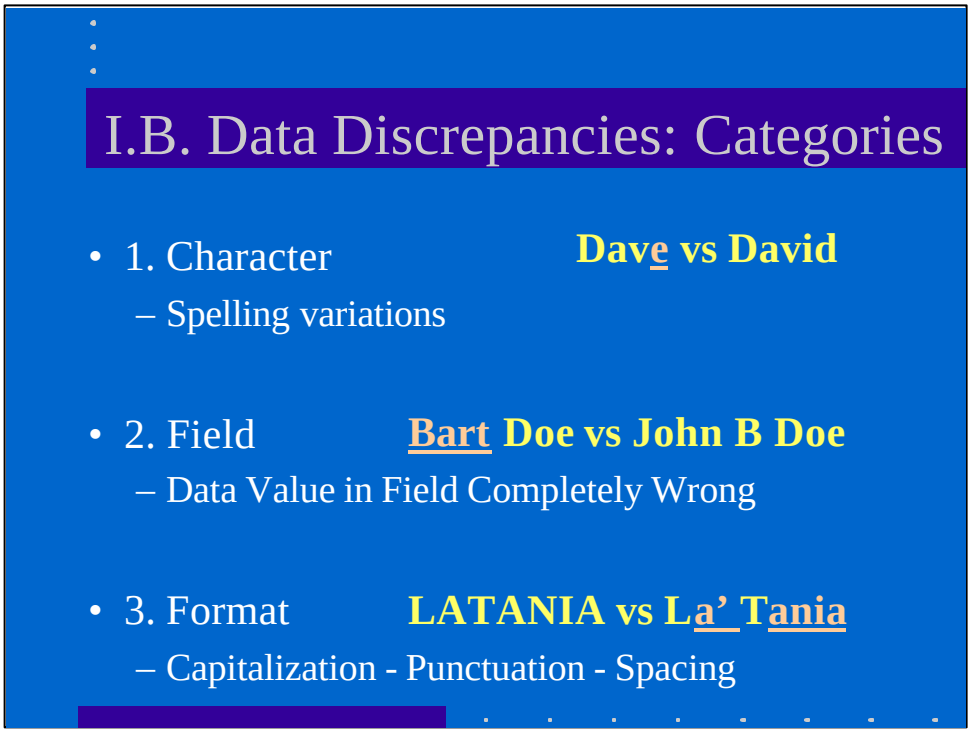
Discrepancies can be due to lack of accurate knowledge about the true value of the data element. In an immunization registry, this most commonly happens when a friend or relative (other than the mother) brings a child to the clinic for a shot. This often leads to a nickname or a variant spelling of the name.

4. Intentionally Disguised Data

This can be done by clinic staff to protect privacy, or by a family for personal reasons.

5. Valid Data Changes

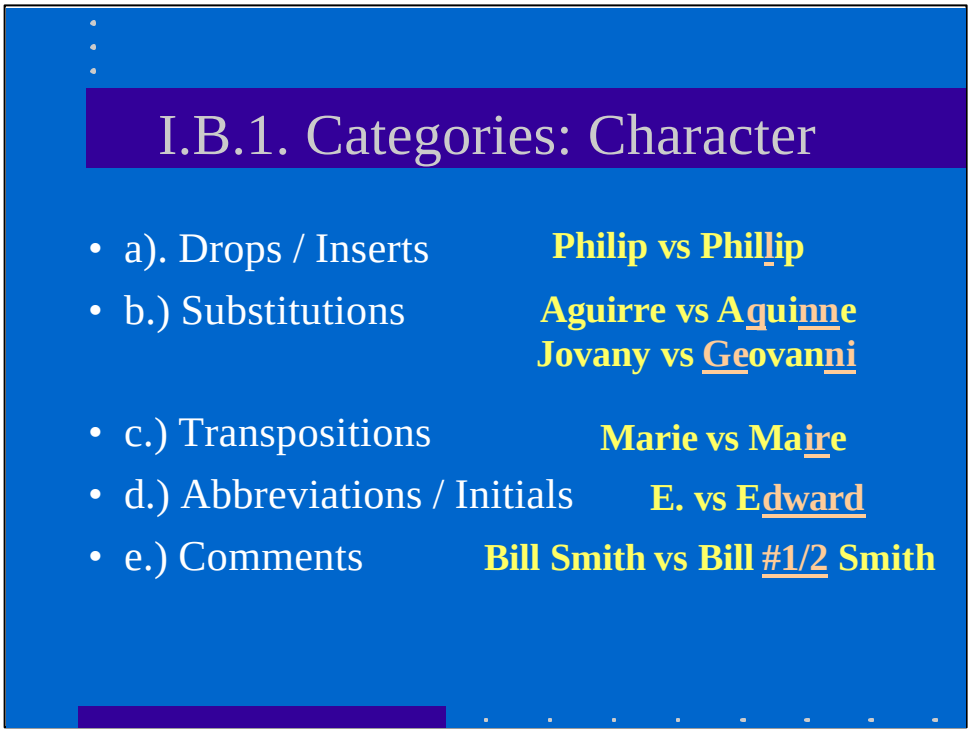
This includes both normal updates to data items such as address and phone number, as well as changes to some data that we think of as being fixed. Data such as first and last names, for example, can be legally changed due to adoption, for religious reasons, or any other reason acceptable to a court of law.



1. Character Errors include all the variations in spelling that can be found in the data, where the field values represent the same underlying data value. They can also include comments added to data, such as Jr. (a generational comment) added into a name field.

2. Field Errors include variations in data caused by entering values into the wrong fields, or due to legitimate changes to data, such as a new legal name, or a new address. This can also include a comment in lieu of a data value, such as "Baby Boy" instead of a first name.

3. Format Errors include variations in data caused by variations in capitalization, spacing, or punctuation. All software which is attempting to recognize discrepancies should be designed to deal with this type of error. Since this is quite easy to do, we don't discuss it further.



a). Drops / Insertions:

Single Character Drops / Insertions. This is the single most common typo, about 31% in our study. Our statistics include only insertions/drops that are internal in the name, rather than a prefix or suffix, and do not include adjacent character doubling, which is broken out separately.

Adjacent Character Doubling. Doubled characters represent about 9% of the typos. Typical cases would include "Philip" vs. "Phillip." This really represents a special case of drops/insertions, but it is useful to categorize it separately, because there are very few twin names that differ only in the presence or absence of a doubled character.

b). Substitution Errors These are the least common typos in our study data, accounting for about 2% of First Name discrepancies in typos. They can be categorized into three groups:

Visual Character Substitution

Phonetic Character(s) Substitution

Keyboard Character Substitution

c). Transpositions All forms of transpositions account for about 5% of typos.

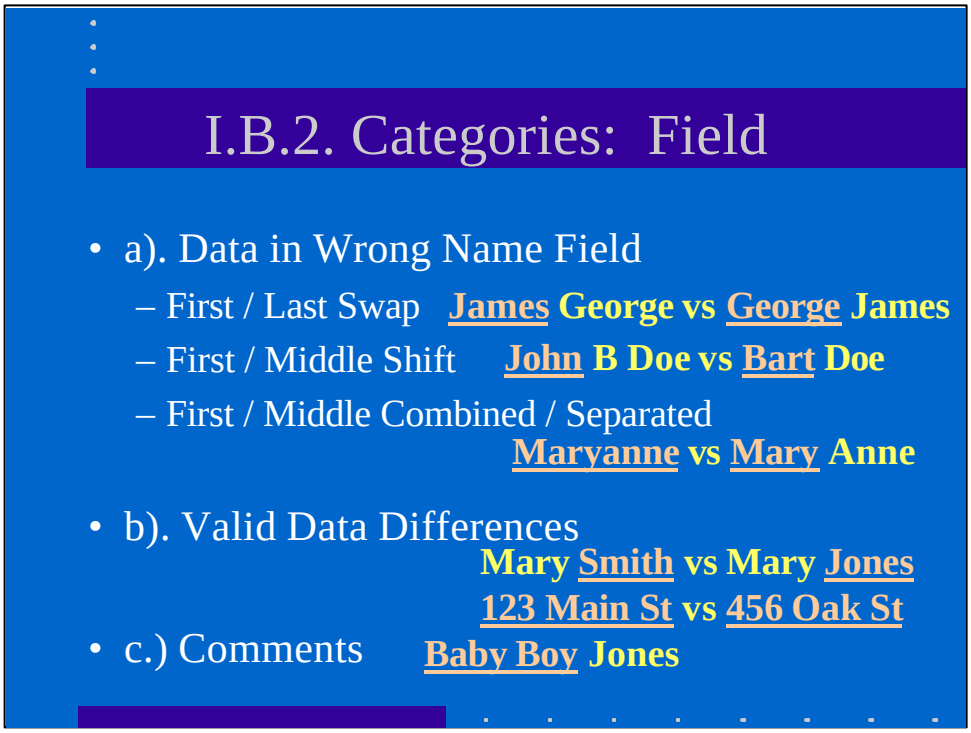
Adjacent Character Transposition

Unbalanced Transposition

Transposition Around Single or Double Pivot Character

d). Abbreviations or Initials Most common in Middle Name (MN) & Address, uncommon in First Name (FN) or Last Name (LN).

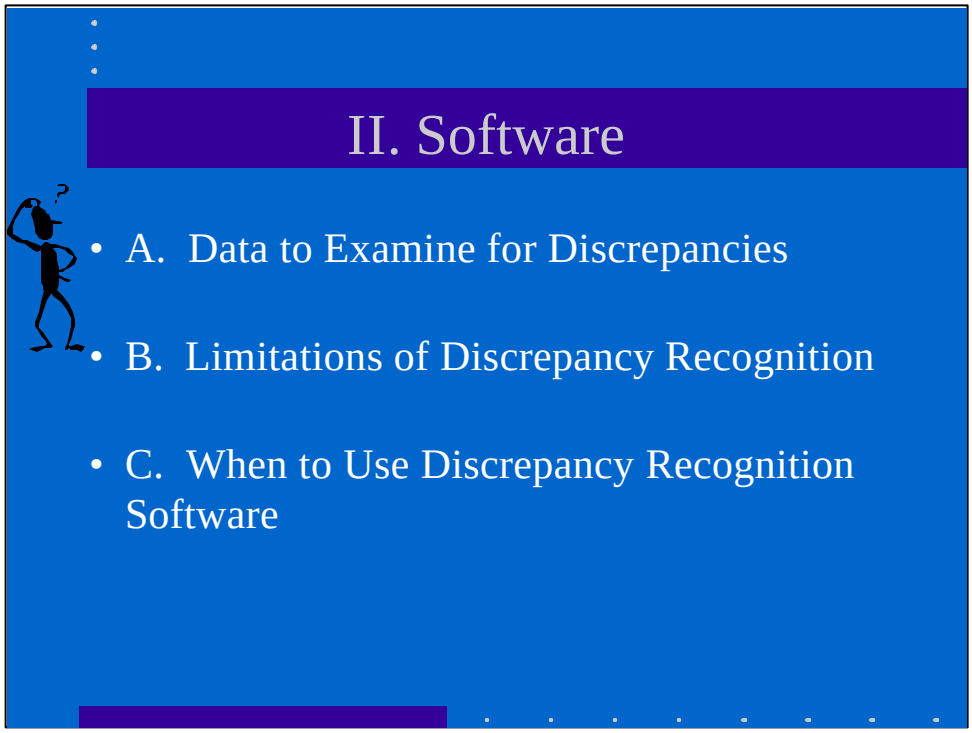
e). Comments Common anywhere in Address and all names**.**



a). Data in the wrong name field is far more common than I ever realized when I first started studying this data. It can account for as much as 30 or 40% of the discrepancies between the names in two records that represent the same person.

b). Valid data differences are far more common in LN, address, and phone number. They are rare in FN and MN (except when the data also ends up in the wrong field). In most cases, there will almost never be a valid difference in DOB, unless the data on the child has been legally changed by a court for the child's protection.

c). Comments sometimes replace the value of an entire data field, such as replacing the address with "Bad Address", or the FN with "Baby Boy" (when the FN isn't known). In other cases, the guardian name may be replaced by something like "Duplicate Use #nnnnn Instead".

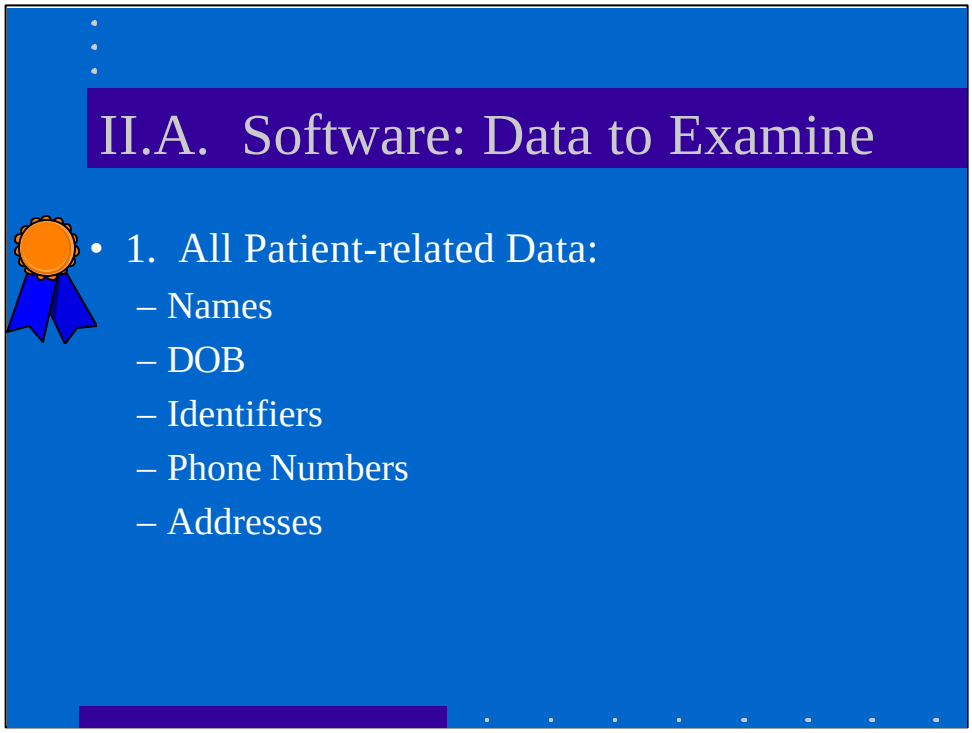


There are three essential issues in creating software to recognize typographical errors:

A. Which fields should we look at?

B. What are the limitations of software which recognizes data discrepancies?

C. When in the de-duplication process should we use discrepancy recognition software?



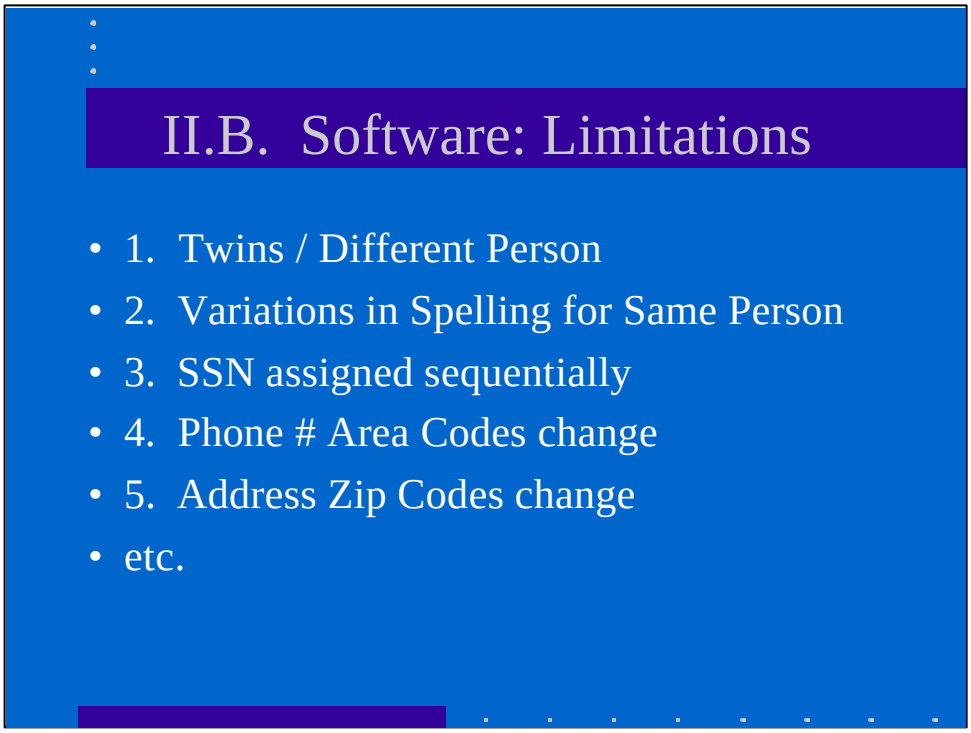
Your software should examine every data element which pertains, even remotely, to identifying the person.

Names of both the client and the parent or guardian are all helpful (including first, middle, and last names).

The **DOB** in an immunization registry is crucial, because recommendation algorithms are based on it. Various studies have shown between 2% and 5% typographical errors in this key identifier, which is lower than any other field we have studied. Most of the DOB typos can be recognized by software.

Unique **identifiers**, including Clinic ID + Chart #, SSN, Medicaid #, Birth Registry #, etc. are all useful.

Phone Number and **Address** are extremely useful. While a difference doesn't rule out two records as representing the same person, a match certainly increases the probability a great deal.



There are a number of situations where software which recognizes typos can lead you to the wrong conclusion. Your software will have to examine these situations carefully, and make appropriate allowances.

1. Twins often have very similar names, sometimes differing by one character, or only in the middle name. For example, “Christian” and “Christina” or “Takeisha” and “Takerisha” occur both as typos and as twins.

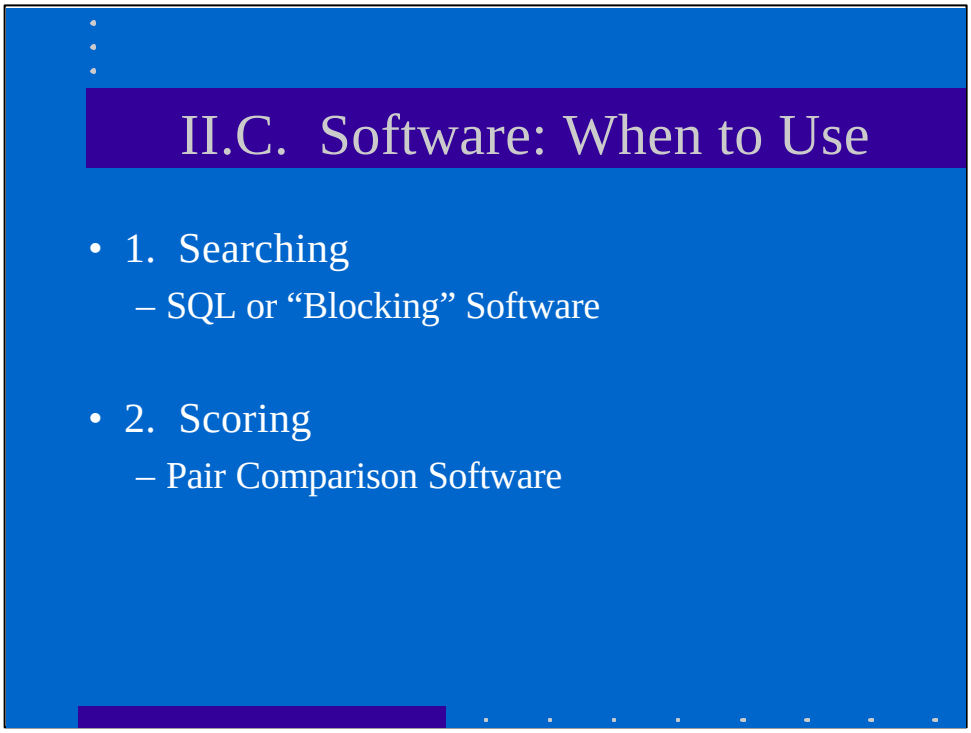
Sometimes even common names can be created by typo transpositions, such as “Mary” and “Myra.”

2. Variations in spelling for the same person can be rather extreme. For example, “Steven,” “Stephen,” “Stefan,” and “Esteban” are all accepted spellings of the same name. It is not uncommon to see a child whose birth certificate says “Esteban” to have his name anglicized to “Steven” by the time he enters the first grade.

3. Even **SSN** have their limitations since they are usually **assigned sequentially**. Since we discovered that one of the most common mistakes in entering a sequence of digits is to be one off in the least significant digit, two persons with similar names and close or consecutive SSN could represent either twins or the same person with a typo in both the FN and the SSN.

Phone Numbers have become less useful over the last decade due to the proliferance of new area codes. You will probably get more mileage comparing phone #s if you ignore the area code.

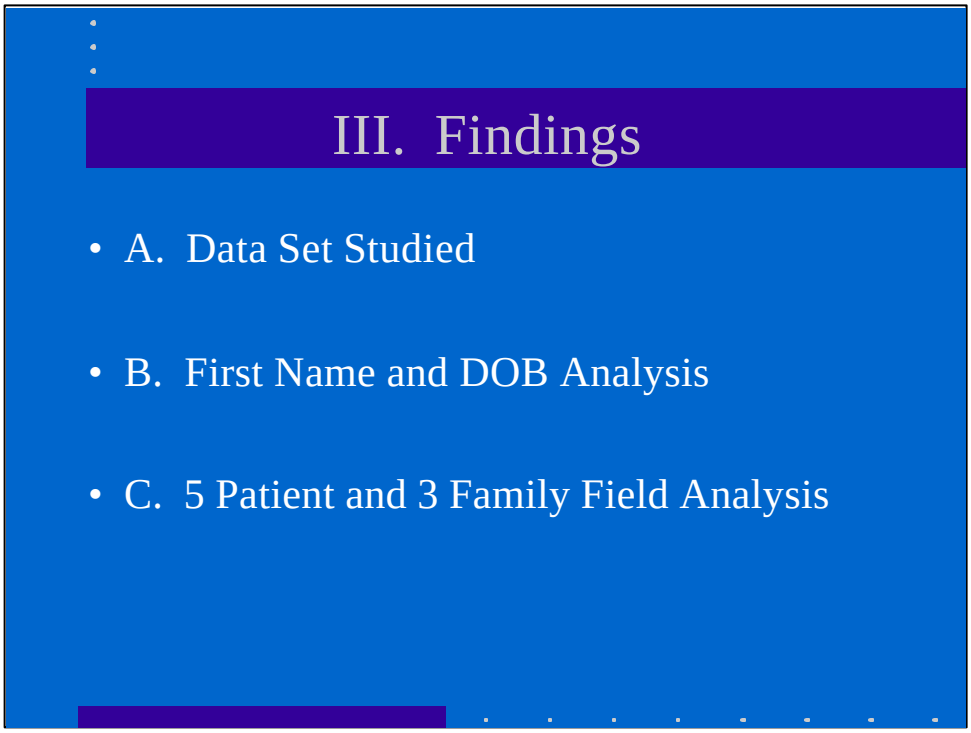
Address comparisons are also subject to the vagaries of changing zip codes, and even changing street names.



1. Searching refers to software techniques that specifically search for common variations in data values as part of the “query” or “blocking” technology. This is most easily done with DOB, because there are a limited number of common variations for any given data value. It is very difficult to do with many other data fields because of the extremely large number of possible discrepancy values. When it is possible to build typo recognition into a query, it can be very productive.

2. Scoring refers to the recognition of typos between a pair of data records. This is usually the easiest place to work on typo recognition.

For more information on the way in which Searching and Scoring fit together into de-duplication software, see http://www.dedup.com for an overview.



Outline of our presentation of findings:

A. The data studied and presented here was gathered from a subset of the Metro Atlanta Team for Child Health (MATCH) Immunization Registry.

It represents the data imported into the Registry in the initial data load of 122,426 records in the summer of 1994, covering children born from January 1, 1988 forward to the date of the initial data load.

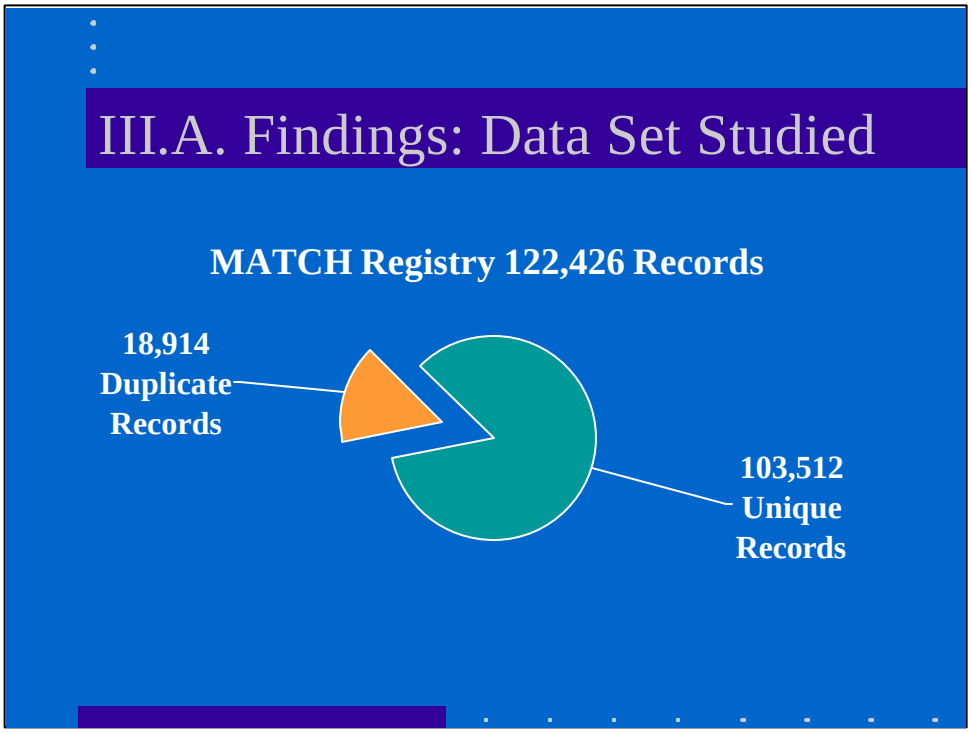
The data came from the two largest counties in GA, and 12 Community Health Centers.

A massive effort, lasting for six months, was undertaken at The Atlanta Project (TAP) to manually de-duplicate the data, with increasing assistance from software during the process.

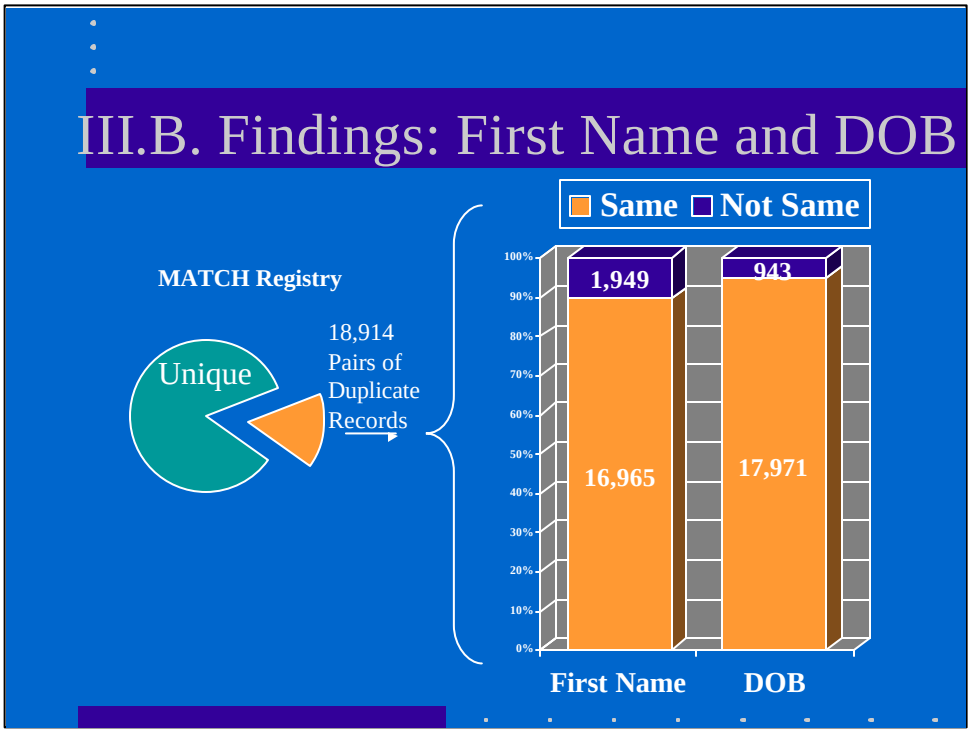
This data was re-examined and automatically de-duplicated by software as part of a CDC sponsored de-duplication study in 1998.

B. First Name and DOB Analysis: These two data elements were studied in great depth, since they are the most basic to the identification of a distinct child, and were present in every record.

C. 5 Patient fields and 3 Family fields were also analyzed for this presentation, in less depth. This analysis provides some results that have strong implications for the viability of Lookups using exact matches on data elements.



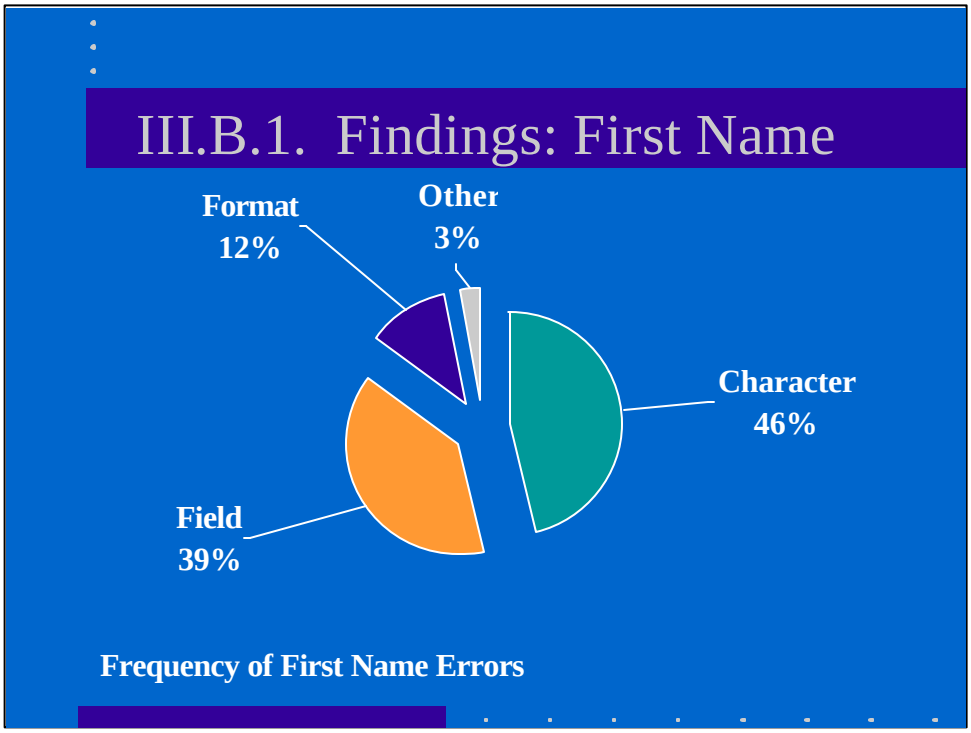
About 15% of this data set were duplicates.



We did detailed analysis of one alpha field, and one date field, because we expect other similar fields to have similar categories of discrepancies. It would have been useful to have also done a similar comparison on a numeric field, such as SSN, but most of our SSN were empty.

About 10% of the duplicates contained discrepancies in the First Name.

About 5% of the duplicates contained discrepancies in the DOB.



From the way we counted the following, there were very few cases where there was more than one specific discrepancy within the major categories. It was common, however, for Field errors to also be combined with Character or Format errors.

Character:

31% - Character Drop or Insertion

8% - Single Doubled Character

5% - Character Transpositions

2% - Character Substitutions

Field:

16% - FN / LN Crossover

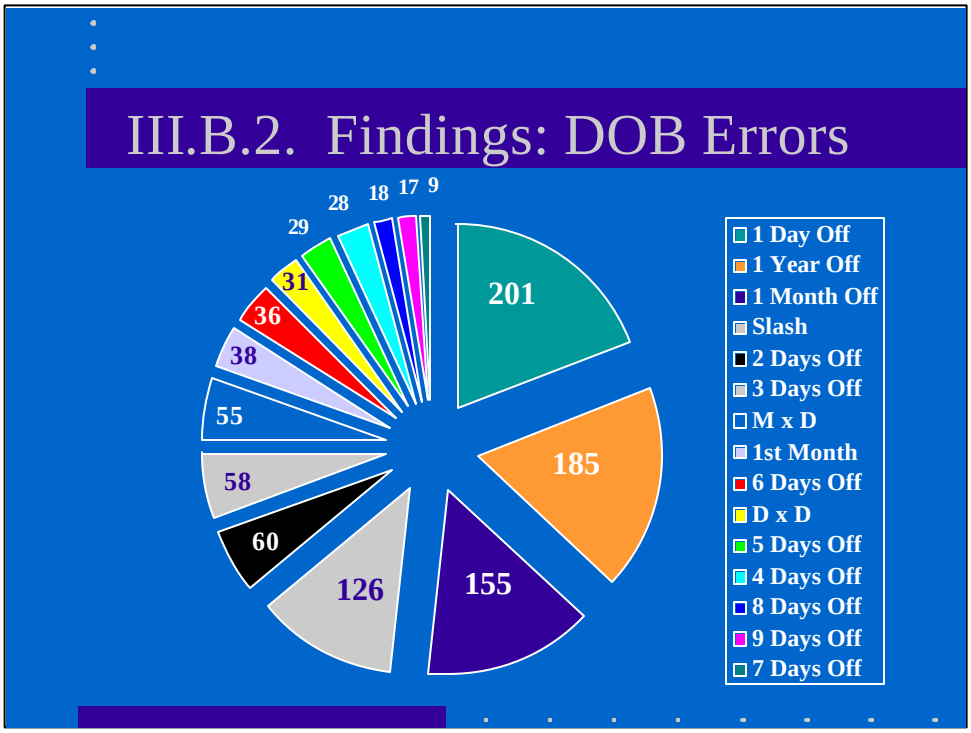
14% - Prefix or Suffix to Name (includes Initials

9% - FN / MN Shift

Format:

12% - Punctuation / spaces / Capitalization

3% - Other



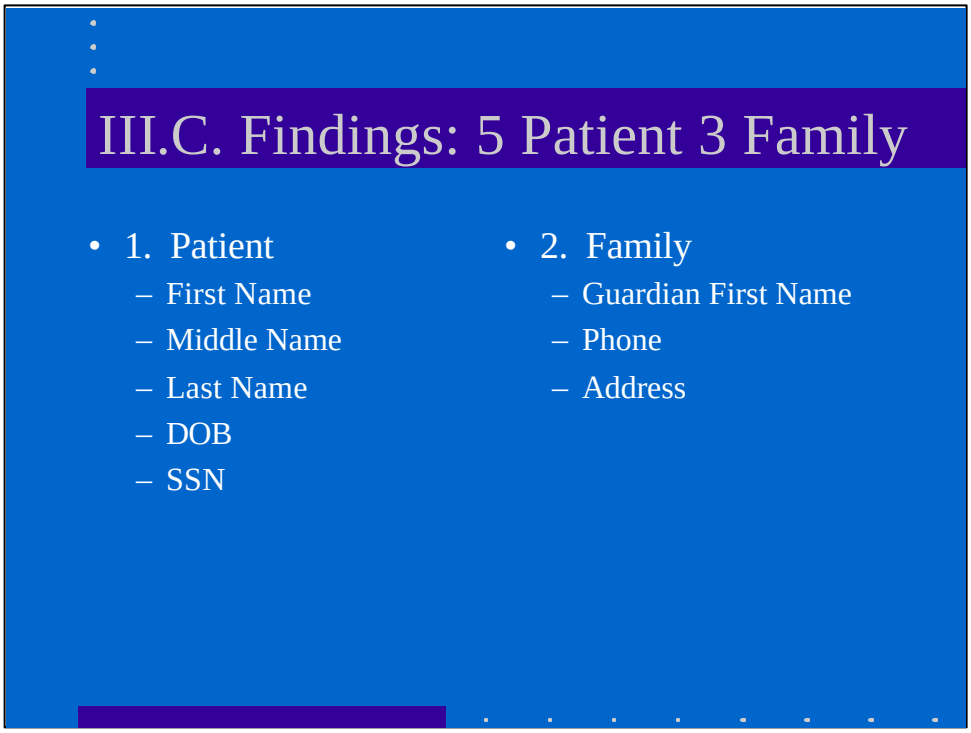
Notes about DOB Errors:

There were 943 Pairs of Records with DOB Discrepancies.

There were 1046 discrepancies identified, because there is some overlap. For example, about half of the day digit transposition errors were also counted as 9 days off.

In this particular study dataset, the only identified errors more than 9 days off were members of the 1 year off, 1 month off, the Slash error group, the M x D group, the D x D group, or the 1st of month group. There may be others that we just didn't find, because the software wasn't looking for them… However, given the declining counts in the "Days Off" categories, it appears that discrepancies of 10 days or more, that didn't fit into one of the other mentioned groups, would be expected to be quite rare.

It is also interesting to note that, in the 541 pairs where one of the date fields is off by 1 (first three categories above), only 142 of them had the same name for parent / guardian. This leads us to suspect that DOB errors are heavily tied to someone other than the mother bringing the child into the clinic for shots.



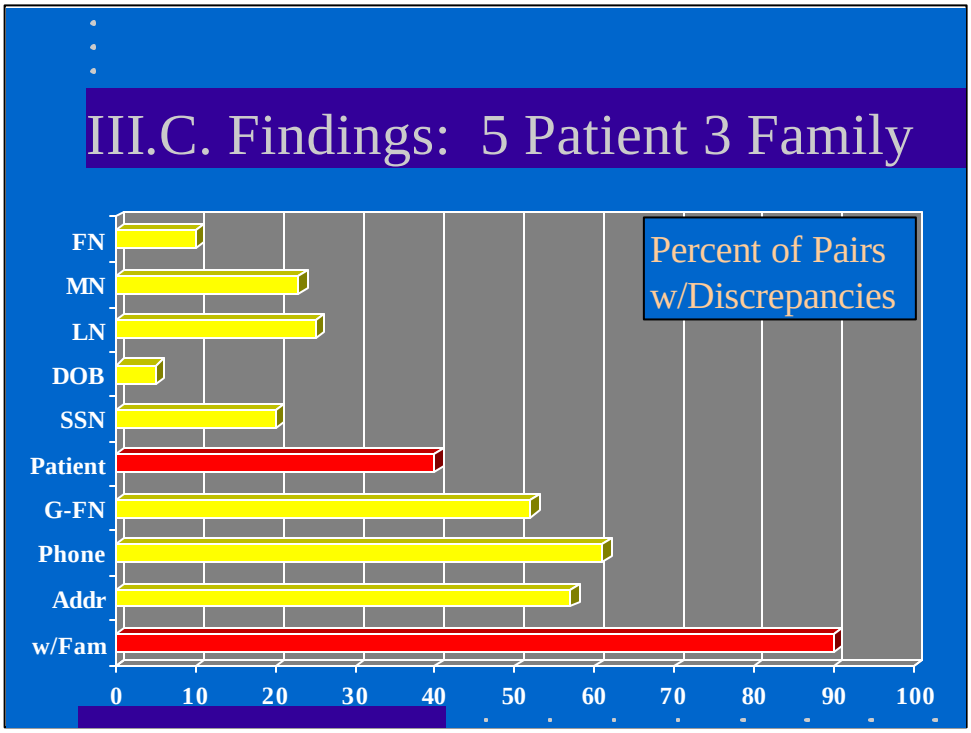
In all of the data elements that follow, we are reporting the number of cases of the total number of record pairs that represent the same person where the value of the data elements are different.

It is important to remember that “your mileage may vary.”

The five data elements on the left “belong” to the patient, and normally don’t change. Most of the discrepancies in this group represent actual data errors.

The three data elements on the right “belong” to the family, and are more likely to change than data values which belong to the patient. In fact, it is common for phone number and address to have a valid data change. It is only slightly less common for there to be a valid indicated change in parent or guardian name, because mothers of newborns may get married or re-married after the child is born, or another person (such as a grandmother or an aunt) may act as guardian if the mother works, and so forth.

Note that lack of data values in fields such as MN or SSN was not called a mismatch.



The yellow bars show the percent of mismatches in each of the 8 data elements studied.

The first red bar shows the percent of mismatches in one or more of the first 5 patient fields.

The second red bar shows the percent of mismatches in one or more of all 8 patient + family fields.

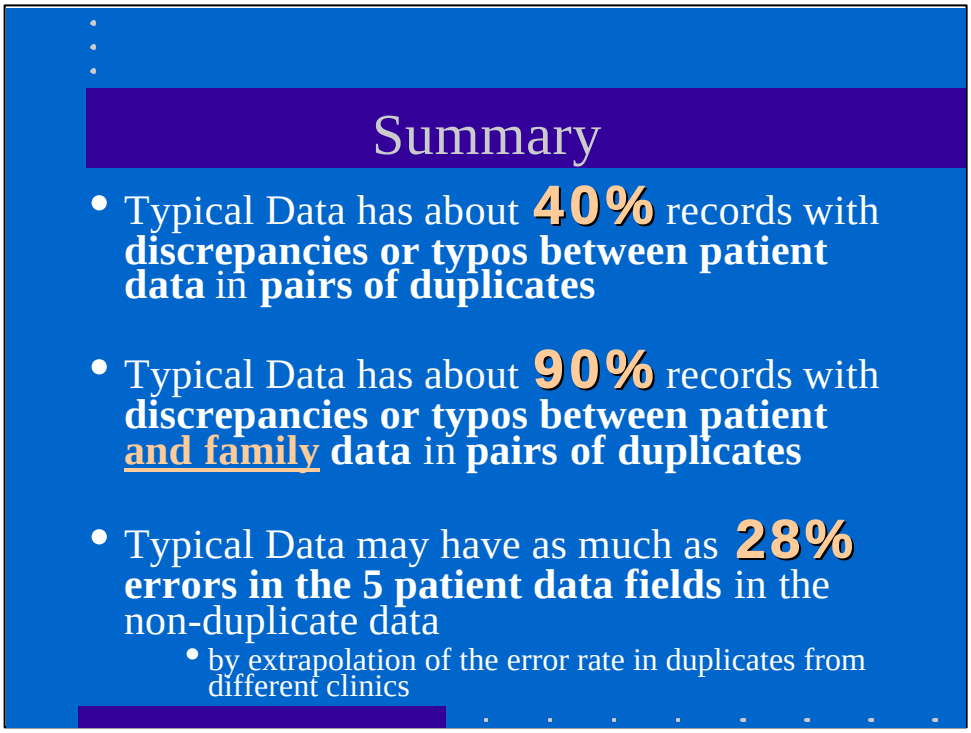
It is interesting to note that **the DOB is the most accurate field** between the pairs of duplicate records. If you consider this carefully, it makes sense. Here are some reasons why we would expect this to be true:

1. Although transposing a two-digit number has a 90% probability of producing a different number, this is not true with dates. If you transpose the digits in the month portion of a date, you only have an **8%** probability of producing a different and valid month.

2. If you transpose the digits in the day portion of a date, you only have a **28%** chance of producing a different and valid day..

3. If you transpose the month and the day values, you only have a **40%** chance of producing valid month values.

4. If you transpose either the day or the month values with the year value, you will **always produce an invalid year** if you collect 4-digit years. 2-digit years will produce an invalid year for dates before 2000.



Calculations:

1. This percentage is from chart on previous page, and represents a count of **7,466** pairs of records with discrepancies in the 5 patient data fields we studied, out of the **18,914** pairs of records which represent the same person.

2. This percentage represents the count of **16,894** pairs of records with discrepancies in one or more of the 8 patient + family data fields that we studied, out of the **18,914** pairs of records which represent the same person.

3. This percentage represents the count of **3,587** pairs of records with discrepancies in the 5 patient data fields we studied, out of **12,630** pairs of records which represent the same person, *but the records come from different clinics.* It seems reasonable to assume that a similar error rate would pertain to all of the data, although there may be some factors which would tend to make the typo rate not that high.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.

Golding, Andrew R. and Dan Roth (1999), *A Winnow-Based Approach to Context Sensitive Spelling Correction*, Mitsubishi Electric Research Laboratory, http://www.merl.com/reports/TR98-07a/.

Hernandez, Mauricio A and Salvatore J. Stolfo (1994) "The Merge/Purge Problem for Large Databases," *ACM / SIGMOD Conference 1995*, 127-138.

Holloway, Geoff (1998), *The Math, Myth, & Magic of Name Search and Matching: Introduction to Name Search and Matching*, Search Software America, privately published, http://www.SearchSoftware.com.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.

John, R. I. And P. R. Innocent (1998), "An Overview of Computational Intelligence," Office for Computational Intelligence, Faculty of Computing Science and Engineering, De Montfort University, Leicester, UK, http://www.csc.dmu.ac.uk/~rij/compiat.html

Knuth, Donald (1973), *The Art of Programming*, vol. 3, Sorting and Searching, Addison Wesley, Reading, Massachussetts.

Kukich, Karen (1992), "Techniques for Automatically Correcting Words in Text," *ACM Computing Surveys*, **vol. 24, no. 4,** 377-439.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press (out of print).

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

Winkler, W. E. (1999), "The State of Record Linkage and Current Research Problems," Census Bureau Research Report, **RR 99/04**, http://www.census.gov/srd/papers/pdf/rr99-04.pdf

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," Proceedings of the Section on Survey Research Methods, American Statistical Association, 467-472
(longer version Research Report **RR 94/05** available at http://www.census.gov/srd/papers/pdf/rr94-05.pdf).

Alvey, Wendy and Bettye Jamerson, ed. (1997), "Record Linkage Techniques – 1997 Proceedings of an International Workshop and Exposition," Federal Committee on Statistical Methodology, Office of Management and Budget. Washington, DC, 1997, National Academy Press, 1999 (not in print).

Notes: This 500 page volume contains numerous articles representing the state of the art in record linkage techniques. It includes one or more articles from most of the people referenced elsewhere in this section. It contains excellent information, if you can find a copy.