

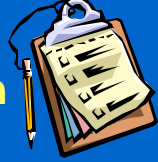
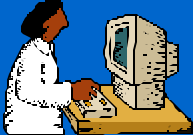
# Designing Data Collection to Avoid Typographical Errors

David Shields, QS Technologies, Inc.  
Cynthia Rust, Emory University

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.

Disclaimer: David Shields is an employee of QS Technologies, Inc. QS Technologies markets software which does de-duplication, as well as shrink-wrapped Immunization Registry, Vital Records, and Public Health Software.

# Outline

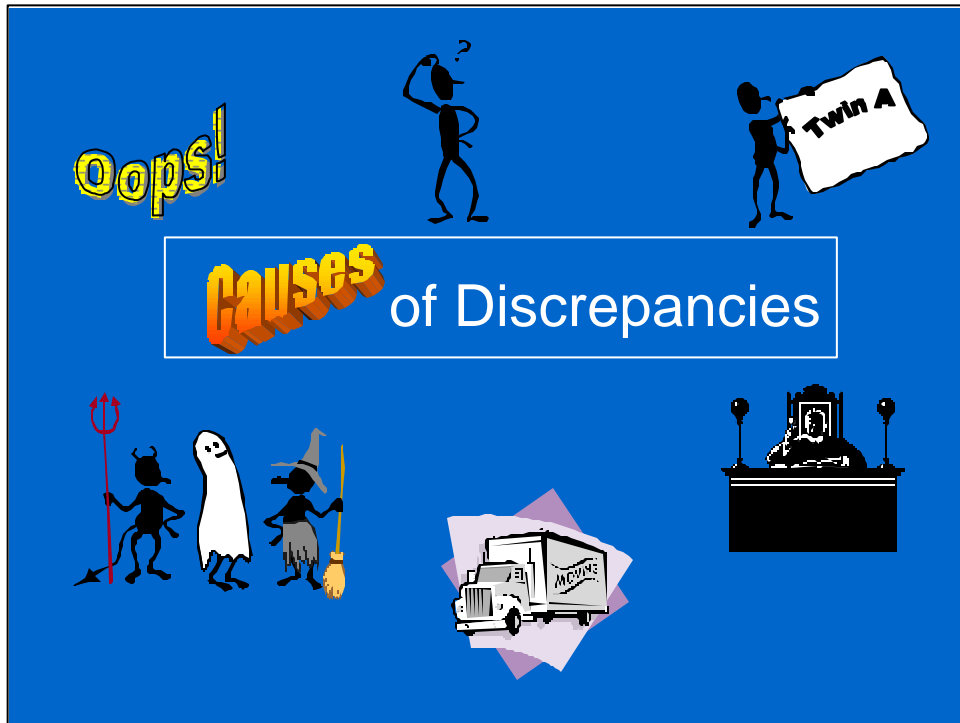
- I. **CAUSES** of Data Errors or Discrepancies
- II. Avoiding Errors in Data Collection 
- III. Avoiding Typos in Data Entry 

2

## Outline

1. The **causes** of Data Discrepancies (which will be called Typos, Errors, or Discrepancies in the rest of this presentation)
2. Some tips on how to avoid errors in initial Paper-based **Data Collection**.
3. Some tips on how to avoid typos in Computer **Data Entry**.

Many of these tips are based on the research that we did into discrepancies between data records when there were duplicates.



### Causes of Typos

**Oops! Mechanical or mental errors** are the basic type of discrepancy that has been traditionally called a “typographical error.” In the narrowest definition, this refers to unintentional mistakes made during the transcription of data.

For the purposes of this presentation, we will also include several other causes of discrepancies under the broad heading of “typos”:

Discrepancies can be due to **lack of 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. For example, fathers are particularly bad about being “not quite right” on the date of birth.


It is very 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.


There are also cases where data is deliberately disguised, either by clinic staff to protect privacy, or by a family for their own reasons.

**Valid data changes** include both normal updates to data items such as address and phone number, when the client moves, 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.

**CAUSES** **Oops!**

# Typographical Errors

- **Data Collection**
  - Confusing & Busy Forms, Tiny Spaces
  - Data Fields not in “Familiar” Order
- **Data Entry**
  - Poor Handwriting 
  - Screen Data Fields not in “Familiar” Order
  - Screen Data Fields not same as Data Collection Forms **Doe, John vs John Doe**

**"My name is NOT Doe, John."** 

4

## Causes of Typos

### **Data Collection**

Forms are

- Confusing
- Overly Busy
- Have tiny spaces to write in
- Data is not in the familiar order that people normally use, such as having the box for entering the family name appear before the box for the given or first name.

### **Data Entry**

Typos increase due to:

- Poor handwriting
- Screen data fields are not in the familiar order that people normally use
- Screen data fields are in a different order or position than the boxes on the data collection forms.

CAUSES

## Lack of Knowledge



- Data Collection

—Person bringing in Patient is not Mother

*"I know he was born on the thirteenth,  
but was it April or May?"*

5

### Causes of Lack of Knowledge

#### **Data Collection**

The single most common cause of lack of knowledge about the true value of data elements is someone other than the mother bringing the patient to the clinic.

Our research has shown that when there is a **discrepancy between the DOB** in one record, and the DOB in another record for the same person, then the **“Guardian” name in the two records is different 80% of the time**. This high discrepancy rate happens with fathers and other relatives, friends, foster parents, and legal guardians.

#### **Data Entry**

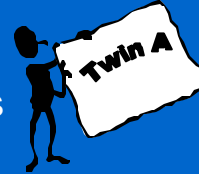
We do not expect our data entry clerks to furnish data about the patient, so this cause of data discrepancy does not apply.

## CAUSES

# Comments in Data Fields

- Data Entry

—Lack of Data Fields for Comments



Name: First  Middle  Family

6

### Causes of Comments in Data Fields

#### **Data Collection**

Comments in data fields on data collection forms are usually recognized as such by the data entry clerk, and are not much of a problem in data collection. Comments such as “Jr.” or “Baby Boy” probably originate on a data collection form, but will be put in the proper place during data entry, if a proper place exists.

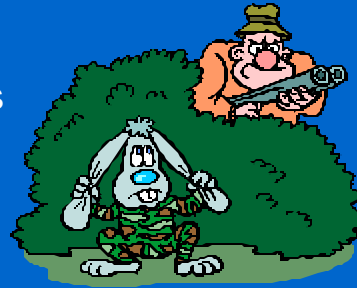
#### **Data Entry**

The single cause of most comments in data fields is the lack of anywhere else for the data entry clerk to put relevant and important information.

## CAUSES

# Intentionally Disguised Data

- Data Collection
  - Parent: Personal Reasons
- Data Entry
  - Clerks: Standing Orders to disguise data from Women's Shelters, etc.



7

### Causes of Intentionally Disguised Data

#### **Data Collection**

Intentionally disguised data will happen during data collection when the parent fears that something “bad” may happen to them if the data is reported accurately.

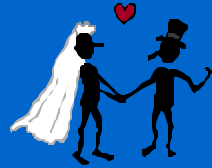
#### **Data Entry**

Data entry clerks may have standing orders to disguise data from Women's Shelters, or HIV clinics, to protect the safety or privacy of the patient.

**CAUSES**

## Valid Data Changes

- **Data Collection**
  - Legal Name Change
  - Foster Care
  - Moved
  - New Phone #
  - Divorce



8

### Causes of Valid Data Changes

#### **Data Collection**

Valid data changes can happen when the parent moves, gets a new telephone number, gets married or divorced. They can also happen when a child is placed in foster care, or given up for adoption. In some cases, a parent may even have a name changed for religious reasons.

#### **Data Entry**


Valid data changes do not normally happen during data entry.



### **Tips on How to Avoid Errors in Data Collection**


We will look at 4 of the 5 groups of Discrepancies that we identified as Causes of Data Discrepancies:

- Typos
- Lack of Knowledge
- Intentionally Disguised Data
- Valid Data Changes



## Errors in Data Collection

**Oops!**

- Avoid Typographical Errors
- Paper Forms must be **Easy to Use**
  - Use a **Professional User Interface Designer**
  - Columnar Format
  - Row Labels on Left
  - Large Boxes to Write In


10

### Tips on How to Avoid Errors in Data Collection

#### **Typographical Errors**

Make sure your paper forms are easy to use. Most “registration” forms at medical facilities make me want to go screaming out of the door. Some of the following ideas may help:

- Consider the use of a **Professional User Interface Designer**.
- Design your forms in **columnar format**, with **row labels on the left**, and very **large boxes** to write in. People whose writing skill is minimal are far more likely to abbreviate, omit, or write illegibly when the boxes are too small.



# Errors in Data Collection

**oops!**

- Data Elements in “Familiar” Order
  - Names: FN - MN - LN
  - Address:
    - Street Address on first line
    - City - State - Zip on second line

11


## Tips on How to Avoid Errors in Data Collection

### **Typographical Errors**

Present data elements on your collection forms in the familiar order that your clients normally use.

**Names should be presented in the way that most of your clients will say them**, which (in the US, anyway) is First (Given) Name, Middle Name, and Last (Family) Name. Since people almost always write their name on one line, your data collection boxes should also be on the same line. This will still leave a little confusion for persons with Asian names, who often give their family name first, but it will reduce errors for everyone else.

**Addresses should also be presented in the way that most of your clients will read and write them**, which is street address on the first line; city, state, and zip code on the second line. Any other arrangement will seem unnatural, and increase the typos made in data collection. If you collect a separate mailing address (PO Box address) from the (physical) street address, it should be on a separate line, following the street address.



# Errors in Data Collection

**oops!**

- Use **simple unambiguous** field labels:
  - “Family Name” instead of “Last Name” or “Surname”
    - Asian names often give “Family Name” first

12



## Tips on How to Avoid Errors in Data Collection

### **Typographical Errors**

Use simple and unambiguous field labels. This is not as easy as it sounds, but here one tip that is generally helpful:

Use “**Family Name**” instead of “Last Name” or “Surname”.

- Some of your clients and data entry clerks may not understand clearly what a “Surname” is.
- Any clients who were born in Asia may think of their “Last Name” as being their personal name, or “Given” name, because it is common with Asian names to say the “Family” name first.
- For the same reason, it may be better to use “**Given Name**” instead of “First Name.”

## Errors in Data Collection

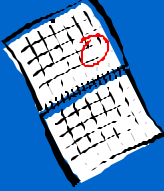

- Dates: **Month - Day - Year**
  - 3 Separate Fields
  - Label fields as **Month, Day, and Year**
    - Many Alternate “Familiar” Orders

02.03.02

^20020302

2 Mar 2002

- Avoid a “/” as a date separator





### Tips on How to Avoid Errors in Data Collection


#### Typographical Errors

Dates have fewer typos than any other field we collect, partly because many typos transform the digits into an invalid date. Nevertheless, there are several important things you can do to reduce the incidence of typos in dates even further:


- Collect the date in the **order that is familiar** for most of your clients.
- Collect the date in **3 separate fields**, rather than one field with separators. This avoids the problem of picking up the first digit of the next subfield. For example, a date of March 2, 2002 can be inadvertently entered as 03-22-002, which some software will parse as March 22, 2002.
- Having the 3 separate date fields then allows you to **clearly label each element** of the date. This helps a great deal when you have any staff or clients who are immigrants from other countries. There are many different standards for ordering the components of the date, and most of the world expresses it differently than we do in the US, typically expressing the day before the month. The only logical standard for expressing the date is the ISO standard, which is YYYYMMDD, but is almost never used in writing dates.
- Avoid a “/” as a date separator**, because it can be read as the digit one. In fact, this is the fourth most common cause of typos in dates, leading to things like January 3, 2002 being keyed as January 31, 2002, or even November 31, 2002.



# Errors in Data Collection



- **Avoid Lack of Knowledge**
  - **ID Cards** Identify Returning Patients
    - Smart Cards
    - Magnetic Cards
    - Bar-Coded Cards
    - Embossing Cards



14


## Tips on How to Avoid Errors in Data Collection

### **Lack of Knowledge**

It is unavoidable that someone other than the mother will bring many of your patients in to the clinic. The mother may have a job, and depend on the child's father, grandmother, or even an aunt or other friend to take the child to the clinic.


However, you can reduce the number of mistakes made when this happens by issuing **ID cards** to your patients. This won't help in all cases, because ID cards are sometimes forgotten, or the wrong one is used when the mother has several children.

Even a paper printed ID card is better than nothing, and increases the likelihood that the child's existing shot record will be found, instead of a new duplicate record being entered under a misspelled name or erroneous date of birth.



# Errors in Data Collection

- Avoid Intentionally Disguised Data
  - Enforce Privacy Policy
  - Separate Medical Data from Financial Data



15

## Tips on How to Avoid Errors in Data Collection

### **Intentionally Disguised Data**

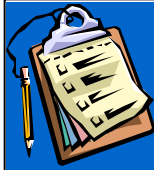
Most data collection forms are completed by the client. Most deliberate disguising of data by the client is **due to fear** that they will suffer something “bad” as a result of accurate data entry. This “something” could be a fear of physical harm, or even the fear of financial harm.

You can help reduce the level of “fear of harm” by attention to implementation of the following:

- Publish an extremely clear **privacy policy**. Teach it to your staff, and enforce it against violations. Then make sure your clients understand that you are serious about restricting access to their medical data.

- Separate** Medical Data and Systems from Financial Data and Systems. This may be difficult to buy into. However, it is clear that someone who has spent time and effort to get an unlisted phone number because they were being harassed by a collection agency, would be very cautious about giving you their unlisted phone number for immunization reminders.

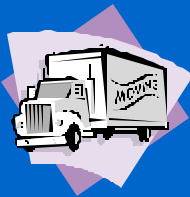
Having worked in a large healthcare agency, I don’t know whether it is possible for the “left hand to be unaware of what the right hand is doing.” However, if you can keep some of the medical contact information separated from your billing department, you will find that it will be more accurate, and will serve its medical purpose better.



## Errors in Data Collection

- Avoid Problems with Valid Data Changes

—Fields for Previous Name, Address, & Phone #



Name	<input type="text"/>
Previous Name	<input type="text"/>
Address	<input type="text"/>
Previous Address	<input type="text"/>
Phone #	<input type="text"/>
Previous Phone #	<input type="text"/>

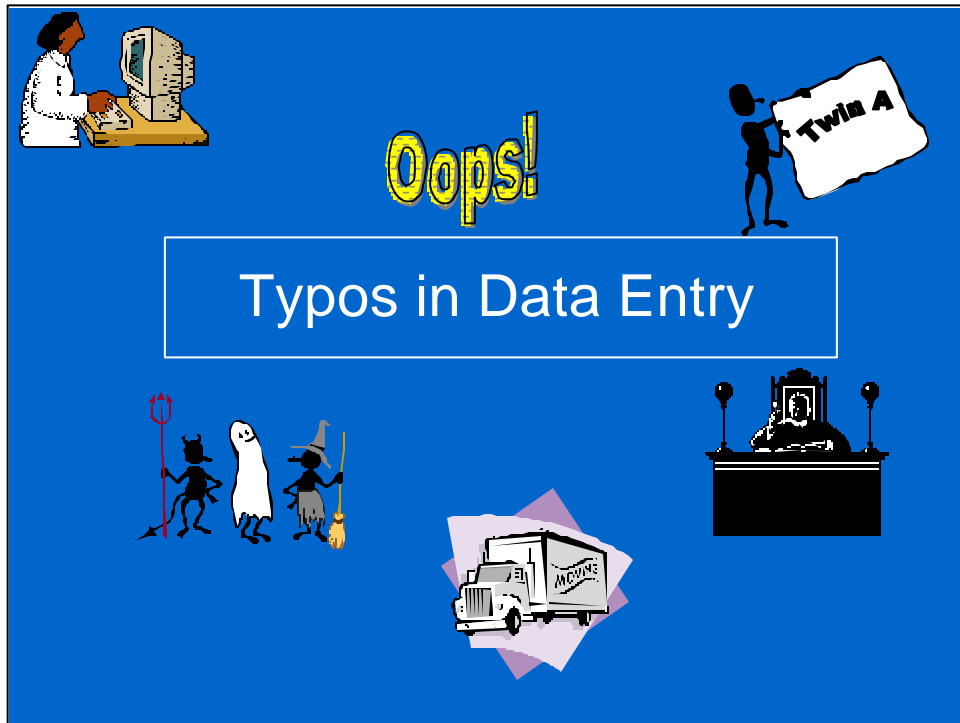
16

### Tips on How to Avoid Errors in Data Collection

#### **Valid Data Changes**

Provide a place on your data collection forms for previous name (sometimes called “Alias” -- although this phrase is best avoided on a data collection form because of its pejorative connotation).


Also provide a place for previous address and telephone #. Explain, on the form, that the purpose of this is to help you locate previous medical records, either at the current clinic, or from another clinic.



### **Tips on How to Avoid Errors in Data Entry**

We will look at 4 of the 5 groups of Discrepancies that we identified as Causes of Data Discrepancies:

- Typos
- Comments in Data Fields
- Intentionally Disguised Data
- Valid Data Changes



## Typos in Data Entry

**Oops!**

- Avoid Typographical Errors
  - Data Collection Forms match Entry Screens
  - Fields in “Familiar” Order
  - Dates in 3 separate fields

DOB: Month  Day  Year


18

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

Some of the same things apply here that applied to Data Collection:

- Make sure the forms your data entry staff is key from match the screens they are keying into.
- Make sure screen fields are in the familiar order that your staff is used to.
- Collect dates in 3 separate and clearly labeled fields.



Oops!

# Typos in Data Entry

- Use Data Entry Verification
  - Two People Key Data
  - Alternatively, Key Data Twice

First Name

First Name, again

19

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

Use Data Entry Verification:

- One possibility is the return to the “keyer” and “verifier” concept used for punch cards. It may be cost prohibitive to do this on every field, but it might make sense to do it on the most important fields, such as DOB and Given Name.
- A second alternative is to have the data entry clerk key the most important fields twice, similar to the way in which we ask users to key a password twice. If they don’t key the same thing both times, then have them verify it again.



Oops!

## Typos in Data Entry

- Validate **Name, DOB, SSN** Data with SSA
  - EVS - Enumeration Verification Service
    - for Employers
  - EVVE - Electronic Verification of Vital Events
    - update SSA records with birth and death records


20

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

This next suggestion is not quite usable at the current time, but could be very valuable to immunization registries, if the proper authorization were obtained from Congress. Either of the following services could be used to verify the accuracy of primary identification data:

- The Social Security Administration (SSA) runs a service called Enumeration Verification Service (**EVS**) that takes Name, DOB, Gender, and SSN data as input, and returns a Yes/No response. A “Yes” response indicates that the SSN is valid for the identified person. A “No” response indicates either that there was a typographical error in the Name, DOB, or SSN, or that the SSN is not valid for the named person. At the current time, this service is provided only to Employers to verify accurate social security withholdings for their employees.
- SSA also runs a second service for State Vital Records Departments, called Electronic Verification of Vital Events (**EVVE**), which is used primarily to update SSA records with birth and death records from State Vital Records Departments.



Oops!

## Typos in Data Entry

- Show **Summary of Data Entry**
  - Display month **alphabetically**
    - shows month - day transpositions
  - Display name fields **together**: FN - MN - LN

21

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

At the completion of a data entry screen, it sometimes help the data entry clerk to spot any typos they made if you **display a summary** of what they entered, arranged in a report format.

- This means that dates would be expressed alphabetically, and names would be displayed together (without large blocks of intervening spaces), the way they would be written in a sentence.
- The same technique would be used for addresses, so that they display in the way they would normally appear on an envelope, without large blocks of intervening spaces between fields.
- Identifiers would also be shown with their normal delimiters, such as nnn-nn-  
nnn for SSN, etc.



Oops!

## Typos in Data Entry

- Do “**Smart Editing**” & Warn if
  - DOB is:
    - more than 3 months before first shot
    - after first shot
    - exactly one month before first shot
    - exactly one day before first shot
  - SSN is wrong age / birth-state
    - use SSA tables for SSN assignment

22

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

Use “**Smart Editing**” and give warnings when improbable data is entered. This can be taken to great lengths, and can help reduce typos. For example, you can do the following:

- Warn if DOB is more than 3 months before first shot (catches 20% one-year DOB errors), or after first shot, or exactly one month before first shot (could be a one-month DOB error), or exactly one day before first shot (could be a one-day DOB error).
- Use the tables provided by the social security administration giving the dates and states where SSN were assigned, and give warning if the SSN is not for someone of the correct age, and the correct state. It is very common for newborns to be temporarily assigned the mother’s SSN for billing purposes, and then the SSN fails to get updated to the SSN belonging to the child until some later point in time.



Oops!

## Typos in Data Entry

- Do “**Smart Editing**” & Warn if
  - Gender is uncommon for Given Name
  - Given Name uncommon and is a recognizable typo to a Common Given Name
  - Given Name is more common as a Family Name

23

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

Use “**Smart Editing**” and give warnings when improbable data is entered. For example, you can do the following:

- It is possible to determine what names have a predominant gender (most of them do), and give warnings when the entered gender doesn’t match the predominant gender.
- It is also possible determine if a given name is uncommon, but is a recognizable typo to a common given name, and give a warning (because the name may simply be typoed).
- It is possible to determine if the given name is more common as a family name, and give a warning. The given name and the family name may be reversed.



Oops!

## Typos in Data Entry

- Do “**Smart Editing**” & Warn if
  - Street Address** elements not recognized
  - Street Address** cannot be validated
  - City Name** does not match **Zip Code**
  - Name-Address-Phone not in **NEWP** (National Electronic White Pages)

24

### Tips on How to Avoid Errors in Data Entry

#### **Typographical Errors**

Use “**Smart Editing**” and give warnings when improbable data is entered. For example, you can do the following:

- Warn if elements of the street address are not recognized. For example, a street address of 123ABC Main Street has alpha characters as part of the street number, which is extremely uncommon, unless the alpha characters actually represent an apartment number.
- Warn if the entire address cannot be validated against a geo-coded valid street address table, which are available for most larger metropolitan areas.
- Warn if the city name does not match the zip-code. Our initial load of the MATCH immunization registry found 127 different ways to spell “Atlanta”, all of which could have been caught and corrected by smart editing at data entry.
- Warn if the Name-Address-Phone# is not in the National Electronic White Pages (NEWP). This may be of limited usefulness, unless the completeness of NEWP improves in the future. Many phone numbers are not a part of this database.



## Typos in Data Entry

- **Avoid Comments in Data Fields**

—Appropriate Data Elements for user comments:

- **Twin / multiple birth / birth order data**
- **Bad Address Flag**
- **Duplicate Record Flag, w/ link to Preferred Record**
- **Given Name Unknown Flag**



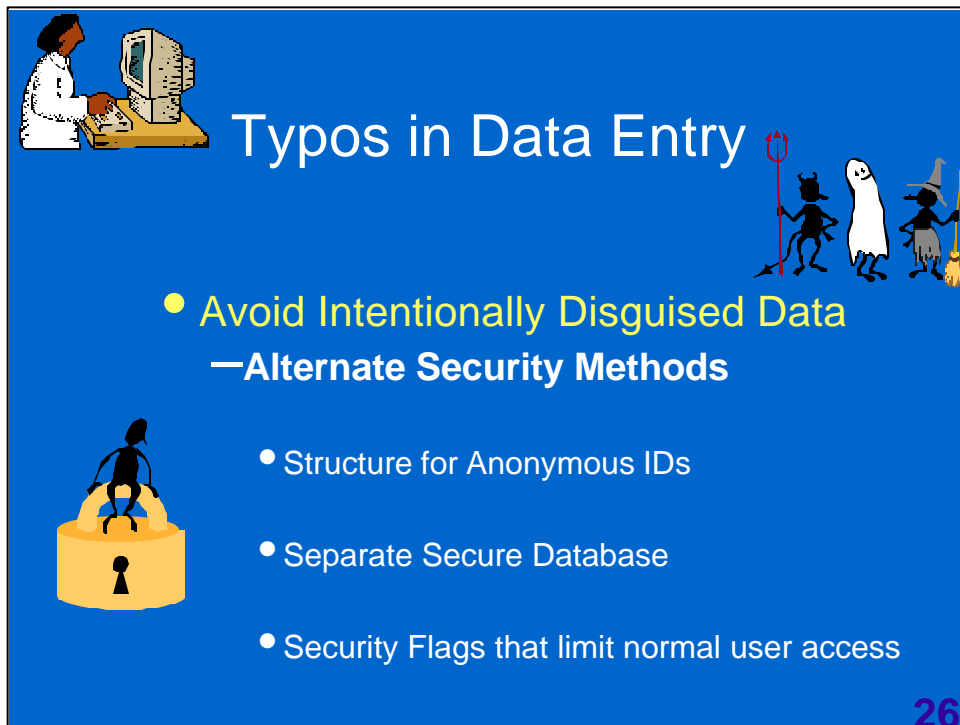
25

### Tips on How to Avoid Errors in Data Entry

#### **Comments in Data Fields**

Provide specific comment fields in your database, and on your data entry screens for the following:

- **Twin / Multiple Birth / Birth Order data.** Distinguishing between twins is the most difficult task for de-duplication and fuzzy lookup software. Accurate use of this field could increase the accuracy of software record matching by an order of magnitude, or more.
- **Bad Address Flag.** Don't make your data entry clerks enter things like "NFA" or "MOVED" in the address field to indicate that a immunization reminder was returned by the post office.
- **Duplicate Record Flag.** When your data quality staff identifies a duplicate record, provide a means for them to flag it as a duplicate, as well as a field in which they can enter the **preferred record** or chart #.
- **Given Name Unknown Flag.** It would be vastly preferable for hospitals and clinics to set a flag to indicate that the child's name is not yet known, than to enter "Baby Boy" or "Twin #1" in the given name field. I have found several dozen variations on combinations of, and misspellings of, "Baby," "Boy," "Girl," "Male," "Female," "Twin," "A," "B," "1," "2," and "#." All of the essential information in these dozens of variations could be captured in 3 fields: the Gender Field, a Birth Order Field, and a Given Name Unknown Flag.



## Typos in Data Entry

- Avoid Intentionally Disguised Data  
—Alternate Security Methods
  - Structure for Anonymous IDs
  - Separate Secure Database
  - Security Flags that limit normal user access


26

### Tips on How to Avoid Errors in Data Entry

#### **Intentionally Disguised Data**

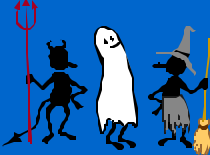
Data Entry Clerks are often taught to disguise data from certain sensitive sources, for the safety or privacy of the client. It would be far better to use one or more of the following approaches in lieu of disguising the data, if it can be done in such a way as to solve the problem:

- Provide a structure for **anonymous IDs**. If you don't provide the structure in the software, the users will create their own, perhaps by declaring that all persons named "John Doe" are an anonymous person. This then creates havoc with de-duplication routines, which may find thousands of persons named John Doe, perhaps all having the same address as well. Since clinics seem to have a need for the creation of anonymous records, set up and design the capability into the software, so that it meets the user needs, and doesn't cause strange results in your reports or de-duplication processes.
- If you absolutely must, you can create a **separate secure database** for confidential data, so that management needs for create statistics on the confidential data are met, and yet individual data is not accessible to normal users. This approach avoids contaminating your report and de-duplication processes.
- A third approach is to set **security flags** in the data records, and make sure that your software respects those flags.



## Typos in Data Entry

- Use **fuzzy search** to locate existing duplicates



27


### Tips on How to Avoid Errors in Data Entry

#### **Intentionally Disguised Data**


One technique to locate data records that have been intentionally disguised by clients is to use a **fuzzy search** through your existing database whenever you enter a new client record.

Our research shows as much as **8% of the records in large county health department databases are duplicates**, caused primarily by typos in DOB, Given Name or Family Name, or by valid changes to the Family Name.

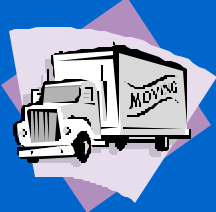
This approach could be used to minimize the build-up of duplicate records in large clinical databases.



# Typos in Data Entry



- **Avoid Problems with Valid Data Changes**
  - Previous Name, Address, & Phone #
    - **Store**
    - **Edit**
    - **Include in Search Algorithms and Screens**



28

## Tips on How to Avoid Errors in Data Entry

### **Valid Data Changes**

You can minimize your problems with Valid Data Changes if you consider previous data values as an important part of your data.

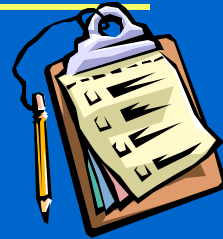
For all fields that have valid reasons to change frequently, you should **store** the previous data value whenever the data value changes.

You should also provide means to **edit or remove** these previous data values (in case an update was inadvertently done to the wrong person's record).

Then, you should use these previous data values in your **search algorithms and screens**. Two people named "Johnny" who were both born on the same day, are probably the same person if one lives at the same address that the other one used to live at.

## Summary: Design To Avoid Data Errors

- **Start** with Data Collection Forms
- **Continue** with Forms matched to Screens
- **Complete** with Screen Design



29

### Summary

Designing Data Collection to avoid errors starts with **good paper form design**, and continues through **coordination of paper forms with computer screens**, and on into **good screen design**.

Ideally the entire data collection and data entry system will be **designed around the goal of ease and accuracy of use**.