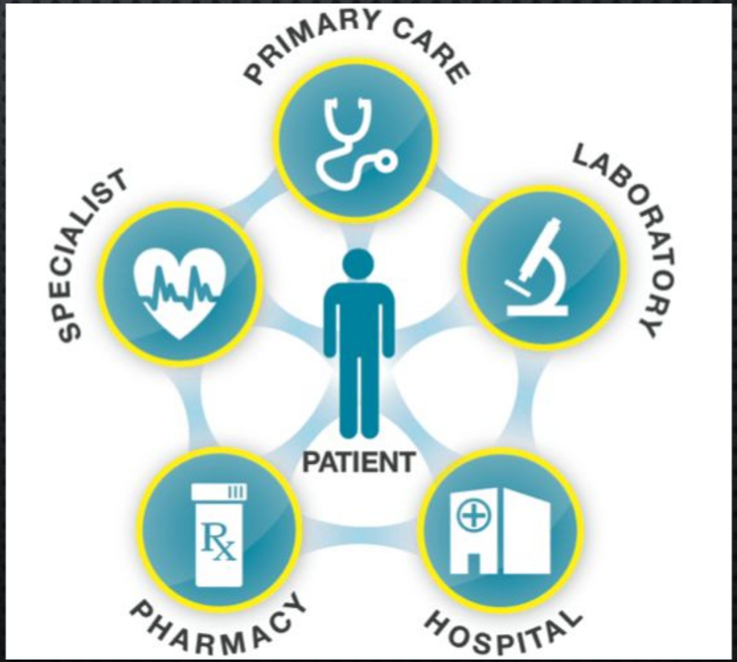
PATIENT MATCHING & DEDUPLICATION

Xinning Chu, Dandan Feng, Kang Fu, Ashley Hall

> “PATIENT MATCHING PROVIDES THE ABILITY TO MATCH A UNIQUE INDIVIDUAL WITH A UNIQUE SET OF DATA IN A HEALTHCARE DATABASE OR DATA SET.”

What is Patient Matching?



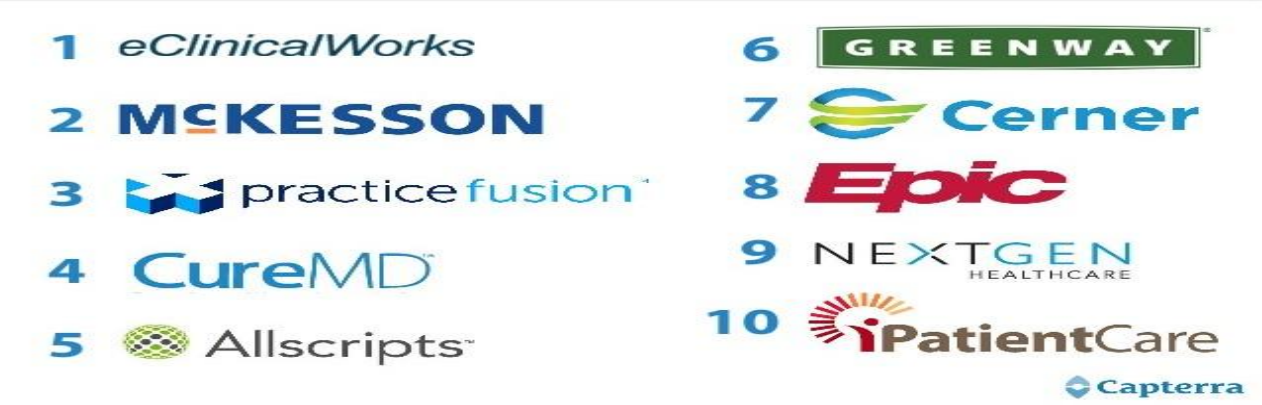
ACCURACY OF PATIENT RECORDS CAN BE SIGNIFICANTLY IMPACTED BY **DUPLICATE RECORDS**. RECORD INACCURACY CAN NEGATIVELY IMPACT PATIENT SAFETY, SPEED OF CARE DELIVERY, AND COST (SPEND).

Safety: An estimated 195,000 deaths occur each year due to medical errors, with 10 out of 17 being the result of identity errors

Spend: Reported averages of $1,950 per inpatient and over $800 per ED visit for repeated medical care.

Delivery: Inaccuracy of patient history due to duplicate records result in repeated tests and delay in ER care and surgery.

Each hospital had dozens of proprietary software vendors, many of them with their own patient ID. Manageable patient matching within a single hospital, and even when consolidation of hospitals happened



1 eClinicalWorks
2 McKESSON
3 practice fusion
4 CureMD
5 Allscripts
6 GREENWAY
7 Cerner
8 Epic
9 NEXTGEN HEALTHCARE
10 iPatientCare

Capterra

HITECH ACT
HEALTH INFORMATION TECHNOLOGY FOR ECONOMIC AND CLINICAL HEALTH

- Signed into law on February 17, 2009 as part of American Recovery and Reinvestment Act of 2009
- Promote the adoption and meaningful use of health information technology

THE ONC PATIENT MATCHING CHALLENGE

PURPOSE :

- Create greater transparency and data on the performance of existing patient matching algorithms
- Spur adoption of performance metrics for patient data matching algorithm vendors
- Positively impact other aspects of patient matching such as deduplication and linking to clinical data

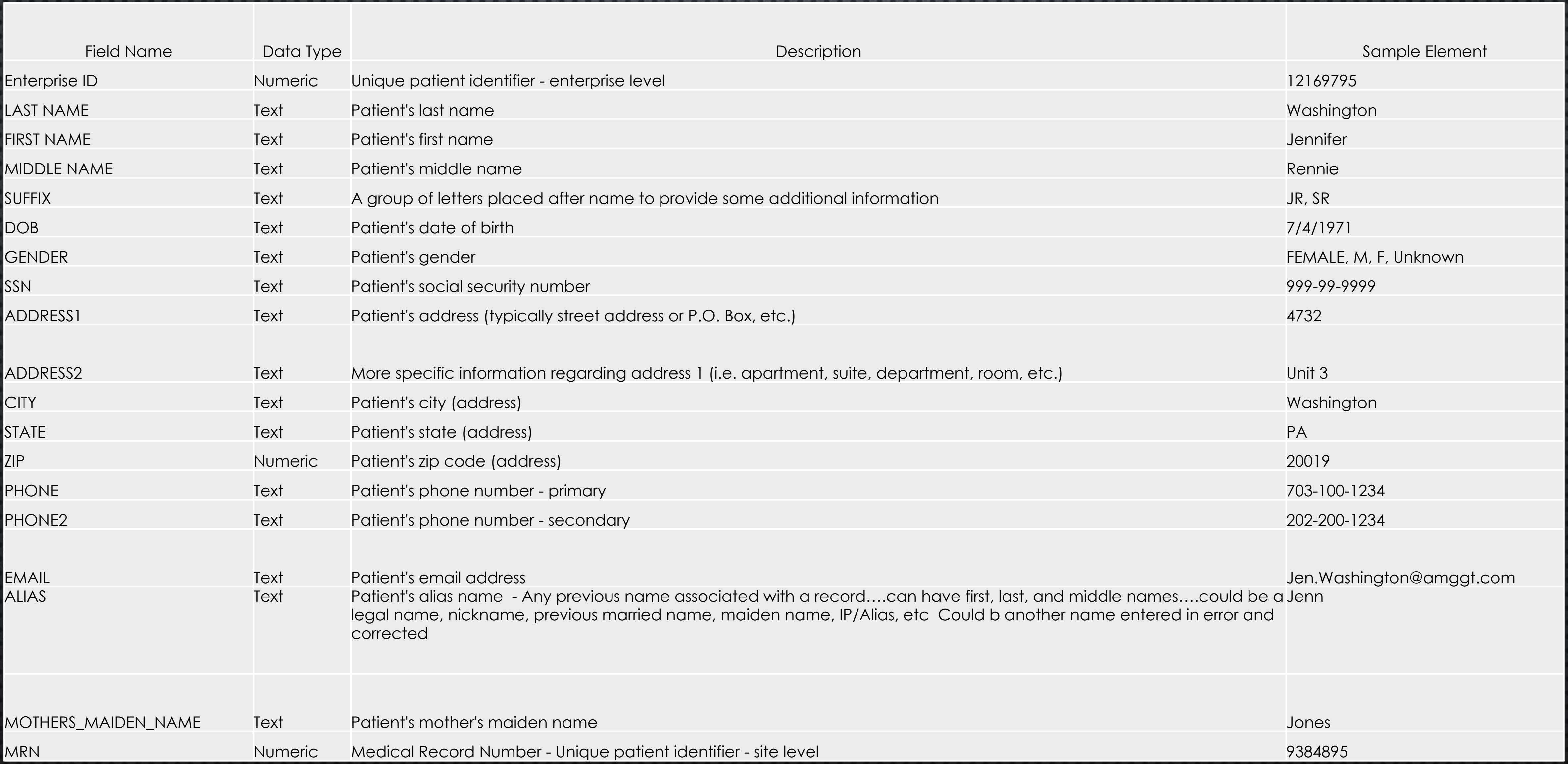
DATA :
Uses a large data set, provided by ONC, against which participants ran deduplication algorithms and provided results for evaluation and accuracy measures. A small set of true-match pairs exist within the large data set, and served as the "answer key".

HOW:
Three Ways to Approach Patient Matching

1. Deterministic Matching: Unique identifiers for each record are compared to determine if two records are duplicates. This method tends to have high precision, low recall, which makes it a strong starting point to become familiar with a data set.

2. Probabilistic Algorithms: The likelihood of duplicate records is determined by calculating the frequency of a value ('John') and the difference between two records ('Jon' vs. 'John'), for example.

3. Machine Learning: A set of rules are created by first "training" an algorithm (various). The algorithm is then applied to the complete dataset to identify duplicate records.



Field Name	Data Type	Description	Sample Element
Enterprise ID	Numeric	Unique patient identifier - enterprise level	12169795
LAST NAME	Text	Patient's last name	Washington
FIRST NAME	Text	Patient's first name	Jennifer
MIDDLE NAME	Text	Patient's middle name	Rennie
SUFFIX	Text	A group of letters placed after name to provide some additional information	JR, SR
DOB	Text	Patient's date of birth	7/4/1971
GENDER	Text	Patient's gender	FEMALE, M, F, Unknown
SSN	Text	Patient's social security number	999-99-9999
ADDRESS1	Text	Patient's address (typically street address or P.O. Box, etc.)	4732
ADDRESS2	Text	More specific information regarding address 1 (i.e. apartment, suite, department, room, etc.)	Unit 3
CITY	Text	Patient's city (address)	Washington
STATE	Text	Patient's state (address)	PA
ZIP	Numeric	Patient's zip code (address)	20019
PHONE	Text	Patient's phone number - primary	703-100-1234
PHONE2	Text	Patient's phone number - secondary	202-200-1234
EMAIL	Text	Patient's email address	Jen.Washington@amggt.com
ALIAS	Text	Patient's alias name - Any previous name associated with a record….can have first, last, and middle names….could be a legal name, nickname, previous married name, maiden name, IP/Alias, etc Could b another name entered in error and corrected	Jenn
MOTHERS_MAIDEN_NAME	Text	Patient's mother's maiden name	Jones
MRN	Numeric	Medical Record Number - Unique patient identifier - site level	9384895

THE DATA – PROVIDED BY ONC

THE DATA – INITIAL EXPLORATORY ANALYSES

- Of the 1,000,000 records 244,606 SSNs records have a null value
- All non-null SSN entries have a character count of 11 (i.e., XXX-XX-XXXX)
- 899 SSN entries are clear dummy entries (i.e., 123-45-6789) but all other SSN's have a maximum count of 3 records
- Genders will likely need to be standardized (M/Male/F/Female/U/Null) and just over 20,000 records are null/U
- 16,463 DOB entries are null
- Most common first, last, and first/last names are to the right. Only 1,417 records have a null value for first and last names.

First	Last	COUNT
ROBERT	SMITH	174
JAMES	SMITH	147
MICHAEL	JOHNSON	132
JOHN	SMITH	130
JAMES	JOHNSON	123
DAVID	SMITH	119
JOHN	JOHNSON	104
ROBERT	JOHNSON	104
JAMES	WILLIAMS	101
MICHAEL	SMITH	100
ROBERT	WILLIAMS	99
WILLIAM	SMITH	98
JOHN	WILLIAMS	94
JAMES	BROWN	91
WILLIAM	JOHNSON	88
MICHAEL	WILLIAMS	88
JAMES	JONES	87
DAVID	JOHNSON	87
JOHN	JONES	86
ROBERT	JONES	86

Last	COUNT
SMITH	7804
JOHNSON	6573
WILLIAMS	5643
BROWN	5114
JONES	4823
DAVIS	3529
MILLER	3490
RODRIGUEZ	3340
GARCIA	2757
WILSON	2740
THOMAS	2591
ANDERSON	2541
MARTINEZ	2516
TAYLOR	2467
JACKSON	2451
MOORE	2393
HERNANDEZ	2226
WHITE	2207
LEE	2179
GONZALEZ	2165

FIRST	COUNT
JAMES	14424
ROBERT	14326
JOHN	14146
MICHAEL	13696
DAVID	11056
WILLIAM	10543
MARY	7754
RICHARD	7634
JOSEPH	7043
THOMAS	6320
CHARLES	6265
MARIA	5646
DANIEL	5323
CHRISTOPHER	5248
PATRICIA	4786
JENNIFER	4526
PAUL	4185
MARK	4168
ELIZABETH	4145
LINDA	4132

THE DATA – CLEANSING IDEAS

- These are potential things that we should do with the data before running anything
  - Enterprise ID
  - Last Name – Remove Special Characters (ie – ‘), all caps
  - First Name – Remove Special Characters (ie – ‘), all caps
  - Middle Name – Remove Special Characters (ie – ‘), all caps
  - Suffix – make sure only values in the dataset currently are SR, II, JR., SR., and JR. So just remove periods
  - DOB -- Could possibly be helpful to break this out into MONTH, DAY & YEAR variables. That way if there is, for example, mistake with the day, the month could maybe still be used in helping with a match.
  - Gender -- FEMALE > F & MALE > M. Make U (unknown values blank?). Do we want to distinguish between genders listed as unknown explicitly and those simply left blank?
  - SSN -- Remove "-“, "Fake' to 'Null'
  - Address 1 -- http://www.gis.co.clay.mn.us/usps.htm has the standard abbreviation for all types of streets according to USPS. We could use this as the reference database for converting to the standardized names, all caps, Do we maybe want to separate address elements similar to how we would separate date elements? Could help control for errors if someone for example entered that an address was a street when it really is an avenure, so there could still be a match highlighted by the number and street name but not the street type (AVE, ST, ETC) part?
  - Address 2 – all caps, same rules as address 1
  - City -- Use zip city, state, & zip code database from USPS to verify different spellings of cities and combinations of zip codes, states, & cities.
  - State -- UN/non-state abbreviates nulled, Use the USPS database to correct any incorrection state abbreivations.
  - Zip -- Add leading zeros to zip codes that are under 5 digits long, Remove any -'s and make all zip codes just 5 digits not 9,
  - Phone 1 -- Make sure all numbers are in 999-999-9999 format,
  - Phone 2 – same as phone 1
  - Email -- Create rules for identifying emails that are clearly incorrect (i.e. they do not include @, do not end in .com .net .mail, etc), all caps
  - Alias – all caps, Thinking we could split the string and make each grouping its own field (i.e. "Lisa Ferguson Potter" into "Lisa" and "Ferguson" and Potter"). From here we can compare to the data entered in First/Middle/Last/Maiden name to either check first name entered or widen the search of that person's last name to maiden name, Some alias's have "^^". We'll have to remove these
  - Mothers Maiden Name -- Remove special characters (i.e. - '), all caps,
  - MRN

THE PATIENT MATCHING APPROACHES

- I was doing a little research on Patient Matching (Record Linkage) and found these resources:
-
- Very interesting article released by NIH on the topic - https://www.ncbi.nlm.nih.gov/books/NBK253312/
- Potential package to use in R - https://r-forge.r-project.org/projects/recordlinkage/
  - Potentially relevant question/answer about the package? - https://stackoverflow.com/questions/36042584/r-recordlinkage-identity
- Interestingly, Google has an interface (no longer updated) for Record Matching - http://openrefine.org/

1. Deterministic Matching: Unique identifiers for each record are compared to determine if two records are duplicates. This method tends to have high precision, low recall, which makes it a strong starting point to become familiar with a data set.

2. Probabilistic Algorithms: The likelihood of duplicate records is determined by calculating the frequency of a value (‘John’) and the difference between two records (‘Jon’ vs. ‘John’), for example.

3. Machine Learning: A set of rules are created by first “training” an algorithm (various). The algorithm is then applied to the complete dataset to identify duplicate records.

- The Health Care Blog – “The Futility of Patient Matching”, Adrian Gropper, MD

- PR Newswire—Black Book – “Improving Provider Interoperability Congruently Increasing Patient Record Error Rates, Black Book Survey”, Black Book Research

- US National Library of Medicine National Institute of Health – “Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields”