Matching Maine's HIE Clinical and APCD Claims Data

Report on Proof of Concept to Match HealthInfoNet Identified HIE Clinical Data to De-Identified Commercial Claims Data From the Maine Health Data Organization

June 5, 2013

Contents

1. Executive Summary

A proof-of-concept was conducted to evaluate the feasibility and effectiveness of matching de-identified commercial claims data from the State's All Payer Claims Database (APCD) maintained by the Maine Health Data Organization (MHDO) to the clinical data in HealthInfoNet's statewide Health Information Exchange (HIE). With grant funding from the Maine Health Access Foundation (MeHAF), and through a contract with the Maine Health Data Organization, HealthInfonet has worked with Arcadia Solutions to test the linkage of the clinical data in the exchange to the APCD. This report provides information on the linkage, identifies problem areas that may need to be addressed to achieve a stronger technical linkage, and discusses the value of the combined data as a more robust database.

The team obtained one year of de-identified commercial claims data from July 1, 2011 through June 1, 2012 and used clinical data from the HIE for the corresponding time period. Claims records contained payer-specific Member IDs, and indirect member identifiers, including date of birth, zip code, gender, and a double-encrypted social security number. The dataset did not include any direct provider identifiers. The team did not expect to be able to match every claim to every clinical encounter in the selected timespan; as the claims dataset was limited to commercial and Medicare Advantage (Medicare coverage provided by Commercial Insurers). In addition, as the HIE is payer agnostic (collecting clinical data directly from the electronic medical record when it is created), it was not possible to identify and eliminate the Medicaid clinical records or the self-pay from the clinical data. Finally, the HIE clinical data contained only those encounters with providers connected to the HIE – at the time of this analysis – 30 hospitals and approximately 250 ambulatory practices. For these reasons, certain events of care on both the clinical and the claims side did not have a match in the opposing dataset.

Specific issues were identified related to the encrypted data. Health Plans submit encrypted social security number or local plan member identifier to the MHDO. The use of these two encrypted IDs by different health plans may have resulted in multiple member IDs if a member had dual coverage or had changed coverage within the time period studied. Also, unique encrypted provider IDs did not allow linkage with the same providers in the clinical database; a step that most likely would have increased the number of matched records significantly. Finally, global claims submitted for hospital services that contained both physician and facility components could not be unbundled and assigned to specific providers.

The primary matching method used date of birth, gender, zip code, and date of service to match claims events of care to clinical events of care and members to persons. While a match on this data combination does not guarantee that the claims and clinical data belong to the same individual, the team found 81% of Member IDs in the claims data and 89% of Person IDs in the clinical data had a unique combination of the demographic data elements in their respective dataset. That is, while it is possible that a Member-Person match based on these data elements is a false positive, using these data elements to rule out non-matches is highly effective. The team then used matches on event of care dates of service and other methods to further refine the matching analysis.

Through this methodology, **264,794 Member IDs (41% of Claim Member IDs)** were matched to **254,120 HIN Persons (33% of Clinical Person IDs)** through demographic information and event of care histories. The team used these matches to develop a provider crosswalk, connecting Provider IDs in clinical and claims data based on matched claims/clinical events of care; this matched **68,352 Claim Provider IDs.**

The accuracy of these results – how many of these matches truly reflect the same person or the same provider – cannot, at this time be confirmed; this report identifies opportunities for validating and improving matching results. While the match rate achieved here may seem low, it is expected that if data from a longer time period were used and identifiers for specific data elements were available; a statistically significant (>95%) match could be achieved across these data sets. For example:

- If the patients that had Medicaid coverage or were self-pay were removed from the clinical data set (due to their having no corresponding claim in the claims dataset)
- If the patients with Medicare and Medicaid coverage were added to the claims data set for matching purposes.
- If provider identifiers were included to allow one-to-one validation of the two data sets at an encounter basis and to support the exclusion of provider claims corresponding to no clinical data due to the lack of HIE connection.

Additionally, the report discusses the potential for using fully identified claims data sets. Once the matching process has been validated and the accuracy rate is acceptable, the matched data could be used to enhance the data currently present in the HIE and the MHDO to both inform providers' care decisions and to support population based analytics.

2. Introduction

2.1. Project Background

HealthInfoNet, Maine's designated statewide HIE organization, is a secure electronic system where health care providers share patient health information including allergies, prescriptions, medical conditions, and lab and test results to better coordinate and improve patient care. Incorporated in 2006 as an independent statewide non-profit organization, HealthInfoNet is one of the leading health information exchange organizations in the country. It is governed by a board of directors and several committees comprised of Maine people serving on behalf of doctors, hospitals, public health, state government, and patients. With strong support and participation from the leading health care stakeholders in Maine, HealthInfoNet has established a true public-private partnership that has achieved the goal of promoting statewide data exchange and use. The clinical data collected on each patient in the HIE provides a broad clinical dataset to promote higher quality and more effective healthcare delivery.

As of March 2013, thirty-four (34) of Maine's 38 acute care hospitals, representing 88% of the state's inpatient and emergency room utilization, 328 ambulatory practices and five Federally Qualified Health Centers are participating in the clinical data exchange. Approximately 1,170,000 patients (87% of all Mainers) are enrolled. This number will grow as more hospitals and provider organizations join the HIE. The exchange includes data on all patients regardless of payment source – insured, uninsured, publicly insured, and underinsured patients are all in the database. Participating providers file data with the exchange on a real-time basis, where it is housed in a statewide data repository organized by a master patient index that links patients across multiple health care settings. Identifying and linking the right patient is a challenging and essential component of the success of the exchange. Finally, HealthInfoNet standardizes the data across sites to guarantee that the statewide data means the same thing to all providers accessing the exchange and that the aggregated database can be analyzed across provider organizations and regions of the state.

In December 2011, HealthInfoNet received a grant from the Maine Health Access Foundation (MeHAF) to establish a data warehouse and demonstrate the potential of linking the clinical data currently in the HIE with claims data from the state's All Payer Claims Database (APCD), maintained by the Maine Health Data Organization (MHDO),[1] as a pilot project.[2] The MHDO is a state agency that collects health care data and makes those data available to researchers, policy makers, and the public while protecting individual privacy. The purpose of the organization is to create and maintain a useful, objective, reliable and comprehensive health information database that is used to

[1] For more information on the Maine Health Data Organization see: http://mhdo.maine.gov/imhdo/

[2] HealthInfoNet (2011). HealthInfoNet Receives Grant to Link Maine's Health Exchange with Statewide Claims Database [Press release]. Retrieved from http://www.hinfonet.org/news-events/news/healthinfonet-receives-grant-link-maine%E2%80%99s-health-exchange-statewide-claims-database

improve the health of Maine citizens. MHDO was also interested in the feasibility of linking administrative claims in the APCD to the clinical claims and as such agreed to contract with HealthInfoNet and provide the restricted data set for this pilot project.

The MHDO receives claims feeds from all commercial payers providing coverage for Maine, Medicaid and Medicare in the State of Maine. Prior to sending claims feeds to MHDO, payers encrypt patient identifying information such as names and social security numbers to ensure the protection of patient privacy. As the resulting claims dataset does not contain direct member identifiers, the pilot project focused on understanding how a fully identified clinical dataset could be linked with a claims dataset without identifiable member information.

2.2. Purpose and Vision

Claims data has been the primary source of payment, utilization, and population health data used for public policy and health services analysis in Maine and across the nation. The MHDO built the first All Payer Claims Database in the country and has been collecting claims data since 2003. The purpose of clinical data, to date, has been primarily focused on informing clinical care decisions and to a lesser extent, research. With the advent of HIEs nationally, the concept of a more robust clinical data set that can inform the evidence base on clinical quality and outcomes of health care utilization has become realized. HealthInfoNet, as one of the most advanced HIEs nationally took on this project to assess the quality of its large clinical data set to be used for analytics purposes and the capability to match it to commercial claims as a pilot to demonstrate the feasibility of linked clinical and claims data that could be used (with appropriate privacy considerations in place) for clinical support (enhancement of the clinical HIE data for clinical decision-making), outcome and cost analysis, and population health research.

Claims data contains information regarding medical encounters for which there is no clinical data in the HIE. This includes all services delivered by providers who are not connected to the HIE; while it is preferable to have the full clinical data on such encounters, claims data contains information that can be used as a proxy for some clinical data and can, at a minimum, indicate that an encounter has taken place. For instance, if a Maine resident has an emergency room visit while traveling out of state, even knowing that the event has taken place enables the primary care provider to reach out to the patient to follow up on the event, ensure that the patient understands and is able to follow discharge instructions, and adjust the patient's care plan as needed. This is particularly important for Maine patients who travel frequently, such as snowbirds, or those who live near the state line and may find it convenient to see providers in New Hampshire or Massachusetts. Additionally, claims data is especially valuable as it presents information on encounters with providers not connected to the HIE: this may include providers with low encounter volumes, or those in specialties such as behavior health that lag in EHR adoption and HIE connectivity.

Beyond providing insight into encounters not represented in the HIE, claims data provides additional information regarding encounters for which clinical data *is* available on the HIE; it provides cost information regarding the encounter. This information allows for the calculation of the total cost of an episode of care or the spending on a patient's care over a given period of time across facilities. In addition, it shows the breakdown of payments by

benefit (amount paid by the insurance, copayment, deductible, etc.), potentially providing insight into cost of care to the patient as well as to the payer.

Due to the breadth of identified clinical data collected by HealthInfoNet since 2009 and the privacy and legal limitations to sharing this data with any entity for purposes other than treatment and operations as defined by HIPAA, HealthInfoNet represented the logical starting place to analyze the feasibility of a linked clinical and claims data set. Beyond this project – a merged dataset combined with strong analytics capabilities could provide a multi-dimensional view of the health of the population of Maine and the care delivered in the state. In addition, a merged dataset offers insight into the quality, cost, and value of care, which may further allow Maine stakeholders who legally have the right to access this data to:

- Leverage a thorough dataset to inform physicians and other care team members in making decisions regarding a patient's care.
- Understand current healthcare spending to inform the designs of payment models, especially in designing systems that reward high quality care and positive patient outcomes rather than services performed.
- Improve population health management by identifying healthcare disparities and delivering population-level interventions in the form of policy changes and private initiatives to address negative trends in disease states and utilization patterns.
- Evaluate the relative outcomes and costs of preventative services and chronic condition treatment plans to identify and promote healthcare delivery models that provide high value, resulting in better outcomes and lower costs across the population.

Maine has the second highest per person medical spending in the country, 24% higher than the national average. While a portion of this spending is due to Maine having an older-than-average population, the high cost of care also indicates an above-average utilization of healthcare services[3]. As such, Maine has the potential to lower health care spending while improving the quality of care; the integration of claims and clinical data creates a dataset supporting such initiatives.

[3] April 10, 2009 ACHSD Cost Driver Report & Recommendations to the Maine Legislature, April 2009

3. Data Sources

This section discusses the claims and clinical datasets used for the matching activities.

3.1. Clinical Dataset

The clinical data used for matching came from HealthInfoNet's Health Information Exchange (HIE), which contains data for approximately 1.1 million persons spanning up to five years of history. The data resides in two separate databases. The first is the Clinical Data Repository (CDR) that houses HL7 transaction messages, received from various hospitals, providers, laboratories, etc. The second is a database that houses patient demographics and the Enterprise Master Patient Index (EMPI).

The EMPI maintains a unique ID for every person in the HIE, which is referred to as the Person ID. For each person, there is a mapping from the HIN ID to each Member Record Number (MRN) that a person acquires when becoming a patient at a medical facility. If a person has visited more than one facility they will have more than one MRN. The availability of a mapping of an MRN to a HIN ID (that is, the mapping of a member to a person) was very important in improving match rates.

For the purpose of the data matching, the team limited this data to encounters that fell within the 12-month span for which claims data based on date of service was available: July 1, 2011 through June 30, 2012. This dataset contained 3.76M encounters representing 763,520 persons and 3,693 providers.

3.2. Claims Dataset

The commercial claims data used for matching was a dataset received from the Maine Health Data Organization (MHDO), and referred to as the All Payer Claims Database (APCD). The dataset contained about 17 million claim lines from the time period July 1, 2011 through June 30, 2012.

Claims Data Elements

MHDO has a number of dataset release types that reveal various degrees of information about the claims:

- General, unrestricted release (fully de-identified data set)
- Restricted release (unrestricted dataset plus limited member information and dates of service)
- Practitioner identifiable release (contains billing provider name)
- Insured group identifiable release (contains policy/group number, which identifies the entity that has purchased the insurance.)
- Restricted and not to be released

The claims dataset provided was designated as the "restricted release" by MHDO. The following diagram shows what data was, and was not, available for the claims.

Claim Line Information	Member Information	Payer Information	Provider Information
Available Data: • Date of Service (Start & End) • CPT & ICD9 Code • Diagnosis Codes (up to 13) • Encrypted Member ID and Provider ID • Claim Status • Facility Type	**Available Data:** • Encrypted Member ID • Date of Birth • City, State, Zip • Gender • Double-encrypted SSN **Not Available:** • Member Name (Encrypted name avail. in red data set) • Social Security Number (Encrypted SSN avail. in red dataset) • Race/Ethnicity (Avail. in purple eligibility dataset) • Group/Policy number identifying purchaser of policy (i.e. employer) (Avail. in orange dataset)	**Available Data:** • MHDO Submitter Code (Prefix + 4 digit number; Prefix indicates gov't vs. commercial vs. TPA. One payer may have more than one code.) **Not Available:** • Payer Name (A single payer may have multiple Submitter Codes)	**Available Data:** • Encrypted Provider ID **Not Available:** • National Provider Identifier (NPI) (Avail. in red dataset) • Provider Last Name or Organization Name (Avail. in green dataset)

Several of these fields posed unique challenges:

Encrypted Member ID: Because the claims dataset was split across two calendar years it presented a greater challenge for matching, due to the fact that many plans operate on the calendar year. The end of a plan year change coincides with members changing plans, and in some cases, this triggers the generation of a new Member ID. In addition, a person may have both a primary and a secondary policy; as a result, a single person may have multiple Member IDs in the MHDO dataset adding the potential for artificial duplication of a unique person in the dataset.

Double-Encrypted Social Security Number (SSN): The MHDO is not permitted by law to release patient-identifying information such as name, SSN, or full zip code[4]. Due to these provisions, prior to submitting data to the MHDO, payers must encrypt the member identifier (ID) field; following this, MHDO again encrypts the field for the "restricted" dataset. While MHDO's encryption methodology is consistent, each payer may encrypt different patient identifiers (member ID or SSN); therefore, when a single person has coverage with two payers, each payer may submit different values for that person's encrypted ID. However, if a person has two policies with one payer in the timespan of the dataset, it is possible (though not guaranteed) that the double-encrypted ID number fields will be equivalent for the two Member IDs.

Encrypted Provider ID: The dataset did not specifically identify providers or facilities. An encrypted Provider ID field did exist; this field was unique to each facility and each payer for each provider. That is, a single physician practicing at two hospitals, each of which bills four payers, would have eight Claim Provider IDs. The field

[4] MHDO can release 3-digit zip codes in the restricted data set.

additionally poses the challenge that global claims are often submitted for hospital services that contain both physician and facility components; this impacts the ability to assign a service to a particular provider.[5]

Claim Statuses

The MHDO dataset had claims in the following statuses:

	Status Code & Description (MC038_STATUS)	Claim Lines
1	Processed as primary	13,885,528 (80.50%)
2	Processed as secondary	2,180,874 (12.64%)
4	Denied	789,468 (4.58%)
19	Processed as primary and forwarded to additional payers	79 (0.0005%)
20	Processed as secondary and forwarded to additional payers	24 (0.0001%)
22	Reversal of Previous Payment	393,021 (2.28%)
	Total Claim Lines:	**17,248,994**

For the purposes of matching, the team used only those claims in statuses 1 and 2 (processed as primary and as secondary). The following information was not used for the following reasons:

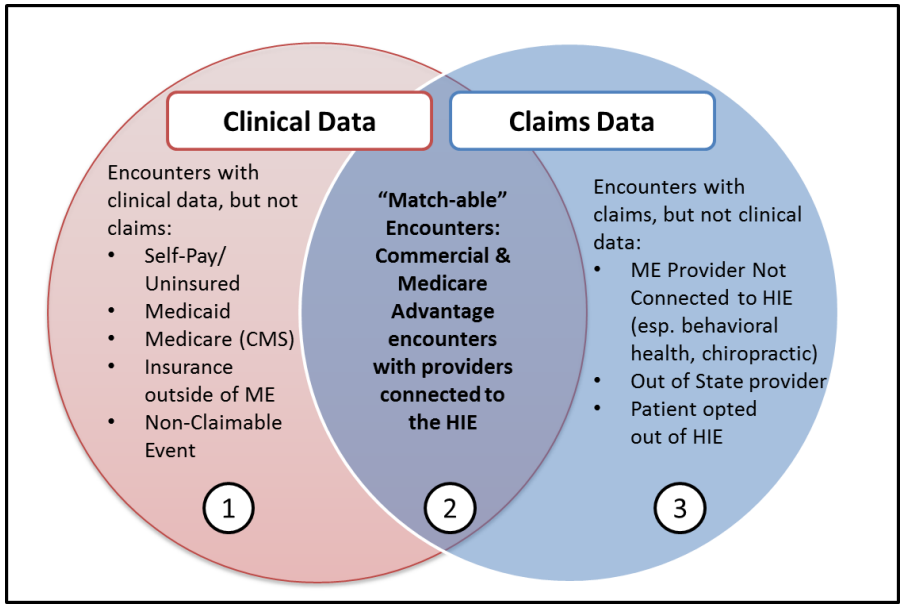
- **Denied Claims.** The reason for this is that denied claims are particularly likely to have low-quality data; one of the top reasons that claims are denied is incorrect patient identified information, such as a date of birth that doesn't match the information the insurer has on file. There are, however, other reasons for a claim being denied, such as lack of coverage for a particular service or documentation such as a referral or medical records not being available. Therefore, excluding all denied claims may also have removed claims that represent valid encounters, resulting in fewer clinical encounters being matched to a claim counterpart. As fields such as date of birth for matching, the introduction of inaccurate data posed risk throughout the experiment that the team deemed sufficient to remove all denied claims from the analysis.
- **Processed and Forwarded.** These claims are ones that would also be reflected in the "Processed as Secondary" status; therefore, these were excluded from the matching to avoid double counting.
- **Reversal of Previous Payment.** This is the reversal of payment made on a previous claim does not represent an event of care and was therefore excluded from the matching.

[5] Maine Health Data Organization "Maine Health Data Organization Data Collection Overview" Presentation to LD 1818 Workgroup by David Vincent, May 10, 2012.
http://www.maine.gov/hit/documents/ld_1818/work_grp_presnt_05102012.pdf

3.3. Clinical and Claims Data Populations

In order to understand the highest match rates possible between clinical encounters and claims, it is first necessary to understand how the clinical encounter data in the HIE relate to claims in the APCD.

A Venn diagram shows there are some clinical encounters for which there are no claims, and some claims for which there are no clinical encounters:



(1) Encounters for which there is clinical data but no corresponding claims data

The data set designated as "1" in the diagram indicates those clinical encounters for which corresponding claims will not exist in the claims data. These data include patients insured by an insurer whose data is not available in the APCD, as in the following situations:

- **Medicaid:** The dataset received from MHDO did not contain any claims submitted by Medicaid (which would have been indicated by a Submitter Code of G0001). While MHDO does collect claims data from DHHS/MaineCare, it was chosen to focus on commercial data for this pilot. Kaiser Health Data for 2011 estimates that about 22% of the population in Maine have Medicaid coverage.[6]

[6] "Health Coverage & Uninsured - Kaiser State Health Facts." Health Coverage & Uninsured - Kaiser State Health Facts. Kaiser Family Foundation. http://www.statehealthfacts.org/comparecat.jsp?cat=3&rgn=21&rgn=1 Accessed 5 March 2013.

- **Medicare Claims from CMS:** The dataset also did not contain any claims submitted by Medicare/CMS (which would have been indicated by a Submitter Code of G0002). The dataset did, however, contain claims associated with Medicare Part A and Medicare Part B products, which were submitted by private insurers.
- **Uninsured/Self-Pay:** Kaiser Health Data for 2011 estimates that about 10% of the Maine population is self-pay[7]; no claims were generated for these events.
- **Patients Insured Outside Maine:** MHDO does not have the statutory authority to collect information on individuals insured outside Maine, such as the following situations:
    - A person who lives in Massachusetts and is insured by Blue Cross Blue Shield of Massachusetts and sees a provider in Maine while vacationing in Bar Harbor. 7% of persons in the HIE have addresses in states other than Maine and are likely to fall into this category.

HealthInfoNet has begun to collect insurance/payment information from some facilities starting in late 2012; as such, this data was not collected for any encounters that occurred during the timeframe for which claims data was available. However, if this experiment were to be done for future claims/clinical datasets, clinical encounters that are not expected to have claims could be eliminated from the matching activities.

Additionally, there are unique situations in which a claim is not generated for care, such as care covered by a grant, such as grants for primary care in severely underserved areas or treatment for pediatric AIDS patients under the Ryan White Act, or care provided as part of a clinical trial. In these cases, encounters may be represented on the HIE without being represented in the claims dataset.

(2) Encounters for which there should be clinical data and claims data

Set "2" is the overlap set that comprises encounters for which there should be data in the APCD as well as the HIE. These encounters meet the following criteria:

- The encounter was with a provider who was connected to the HIE at the time of the encounter
- The encounter was submitted as a claim to an insurer whose data exists in the APCD set
- The patient did not opt out of the HIE.

(3) Encounters for which there should be claims data but not clinical data

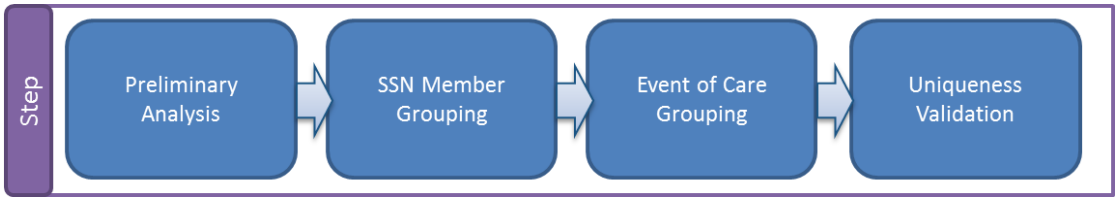
Set "3" in the diagram are the claims in the APCD that have no corresponding encounter data in the HIE. This includes:

- Encounters with out-of-state providers, such as the following situation:

[7] "Health Coverage & Uninsured - Kaiser State Health Facts." Health Coverage & Uninsured - Kaiser State Health Facts. Kaiser Family Foundation. http://www.statehealthfacts.org/comparecat.jsp?cat=3&rgn=21&rgn=1 Accessed 5 March 2013.

  - A person who lives in Maine travelling to another state and seeing a provider.
  - A person who lives outside Maine but has insurance through a Maine provider, such as a college student in Boston covered under his Maine parents' insurance, or a person who lives in New Hampshire but is employed by a Maine company that uses Maine insurance. These situations appear to be rare; only 0.03% of the members in the MHDO dataset reside in a state other than Maine.
- Encounters with providers in Maine not connected to the HIE; in particular, there are specialties that are underrepresented on the HIE due to lagging EHR adoption, such as behavioral health, long-term care and chiropractic care.
- Encounters involving patients who opted out of the HIE.

4. Data Preparation



To prepare the data for matching, the team took the following steps:

- Preliminary analysis of key data elements
- Grouping Member IDs by Double Encrypted Social Security Number
- Grouping Claims and Clinical Encounters, respectively, by date of service
- Reviewing the "uniqueness" of data elements to be used for matching.

These steps are described below.

4.1. Preliminary Analysis of Key Data Elements

The team reviewed the most frequently occurring values in key fields in claims and clinical data to review the quality of the data and to note if there were any fields where particular values were overly represented to the point where the data would not be useful for matching.

The findings were as follows:

Date of Service did not reveal any significant irregularities.

Zip Code also did not reveal any irregularities. Given the varying population levels in different areas, zip code values were not expected to be evenly distributed through the datasets, but the distributions were expected to be at least somewhat consistent between the claims and clinical datasets; of the ten most common values in each data set, nine could be found in the other dataset, indicating that neither dataset was overwhelmingly skewed towards a particular geographic distribution.

Date of Birth also did not reveal any significant irregularities.

January 1st (of various years) was by far the most frequent date of birth in the HIE. The date of January 1st (for all years) occurred for 3,551 people; the average occurrence frequency for all other dates was 2,086 (February 29th was excluded from this average; this date occurred 520 times, or, predictably, 25% as frequently as other days). This suggests approximately 1,465 individuals whose date of birth may be inaccurately documented or require additional review, as January 1. There may be many causes for this. It is known that some persons of Somali ethnicity and recent emigration status have birth dates listed as January 1st as there is no concept of birth-date in their culture. Other causes may be due to the registration event and data entry into the EHR system when the exact date was either not known (it may, for example, have been illegible on the

patient's paperwork) or not entered into the EHR. As date of birth is used for matching, this may have an impact on match rates, as a person may have an accurate date of birth in claims but an inaccurate date in their clinical information. As a result, members with a "true" date of birth of January 1 may find false positive matches in the clinical data, and members whose dates of birth are inaccurately documented as January 1 in the HIE create false negatives.

Primary Diagnosis and Procedure (CPT) Codes: Given the varying prevalence of certain disease states and procedures over others, the team did not expect diagnosis and procedure code values to be evenly distributed throughout the dataset. The following findings were relevant:

- **Completeness:** The primary diagnosis was filled in consistently on Clinical and Claims data. By contrast, the CPT code was null for 12.4% of encounters and 6.68% of claim lines (making "NULL" the most common value in clinical data and the second most common in claims data).
- **Consistency in Diagnosis Codes:** Six diagnoses were among the ten most frequent diagnoses in both the claims and clinical sets; these were diagnosis codes indicating hypertension, diabetes, hyperlipidemia, mammograms, and routine general medical exams for adults and infants/children. Of the four most common codes that were not represented in the clinical "top ten," three were related to back pain (two for non-allopathic lesions, one for lumbago (lower back pain), demonstrating the known underrepresentation of chiropractic care encounters in the HIE.
- **Consistency in CPT Codes:** In addition to "NULL", five other procedures were among the ten most frequent procedures in both the claims and clinical sets; these were two CPT codes indicating an office or other outpatient visit for the evaluation and management of an established patient, and three related to blood work (collection of blood, blood count, and comprehensive metabolic panel).

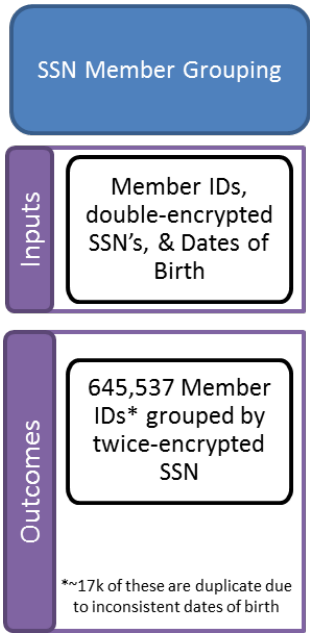
Two discrepancies in CPT codes point to the differences in representation of certain service types:

- The four most common claims procedures, which were not frequent in the clinical data, were related to physical therapy, chiropractic care, and behavioral health, specialties that are not represented in the HIE today due to their low participation rates discussed above.
- Two codes representing Emergency Department visits, 99283 and 99284, represent 5.23% of clinical groups and 1.07% of claim groups, suggesting that Emergency Department visits are underrepresented in the claims data; this is reasonable, as Medicaid and traditional Medicare (CMS) claims were not present in the dataset, and self-pay patient encounters do not generate claims. These are groups that use Emergency Department services significantly more than those with commercial insurance; one study found that nearly one-third (32%) of Medicaid enrollees used the ER at least once during a 12-month period in 2007, while individuals with private health coverage were only about half as likely (17%) to visit an ER.[8]

[8] Garcia TC, Bernstein AB, Bush MA. Emergency department visitors and visits: Who used the emergency room in 2007? NCHS data brief, no 38. Hyattsville, MD: National Center for Health Statistics. 2010.

The detailed results of this analysis can be found in **Appendix A**.

4.2. Double-Encrypted Social Security Number, Date of Birth, and Member Count



Double Encrypted Social Security Number

The Double Encrypted Social Security Number field was populated for 66% of Member IDs. The data contained 417,933 distinct Double Encrypted Social Security Numbers. In addition, 211,329 Member IDs did not have a Double Encrypted Social Security Number. This suggested that there were no more than 629,262 distinct individuals represented by the data.

The team used the Double Encrypted Social Security Number to group Member IDs that belonged to the same person. In situations where two Member IDs had equal Double Encrypted Social Security Numbers, the two were grouped together into a single Member for matching purposes.

Date of Birth

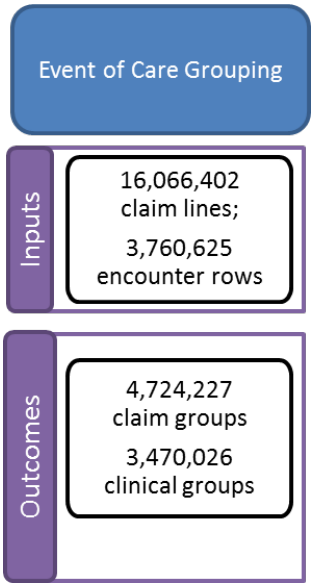
The team found approximately 1,530 Member IDs that had inconsistent dates of birth across the claim records. Member IDs with multiple dates of birth will manifest themselves as multiple records, with one record for each date of birth.

Resulting Member Count

This resulted in 645,537 Members that will be used for matching.

4.3. Event of Care Grouping

The first issue that the team sought to address was the lack of a one-to-one relationship between a claim line (the core unit of data in the claims dataset) and clinical encounter. Due to the manner in which providers bill for their services, a single encounter on the clinical side may generate a multitude of claim lines. Similarly, multiple clinical encounters can correspond to a single claim. To address this, the team combined data points on the claims and clinical sides into event of care groupings that would have a one-to-one grouping.



Method

In the claims dataset, the team defined a **claim group** as the set of all claims associated with a single Member ID, or multiple Member IDs with a single double-encrypted SSN, with the same Date of Service (the field FDATE). In situations where a single claim spanned multiple days, the first date was used. The team used a similar process for clinical data, grouping all encounters associated with a single HIN ID with the same Date of Service into an **encounter group**.

The following scenarios demonstrate how grouping works:

Scenario	Claims Grouping Result	Clinical Grouping Result
Two people who happen to have the same date of birth, gender, and zip (say two twin sisters who are neighbors) have doctors' appointments on the same day.	As the two people have distinct Member IDs, two distinct claim groups are created.	As the two people have distinct HIN Person IDs, two distinct encounter groups are created.
A single patient with primary and secondary insurances has a doctor's appointment; claims are submitted to both insurers.	There is no way of knowing that the two Member IDs refer to a single person; as a result, the two claims are not connected, and two claim groups are created. However, if the patient's policies are with the same insurer, it is possible that the double-encrypted SSN would be the same for both policies; in this case, one group is created.	As the patient has a single HIN Person ID, one encounter group is created.

Scenario	Claims Grouping Result	Clinical Grouping Result
A single patient undergoes several procedures in one day, all of which are billed to a single insurance policy.	Regardless of whether the procedures are billed on one or multiple claims, as long as the date of service and Member ID are consistent, one claim group will be generated.	As the patient has a single HIN Person ID and the events took place in a single day, one encounter group is created.
A single patient is admitted to a hospital for three days, during which time he/she is seen by multiple specialists.	If a single claim is submitted for all events, a single claims group will be created with the date of service. Where multiple claims are submitted for different days, it is likely that there will be as many as three claims groups (one for each day).	Depending on how the care is logged in the EHR, there may be one encounter group or as many as three (one for each day).
A single patient undergoes several procedures over two consecutive days, all of which are billed to a single insurance policy.	If the procedures are billed on a single claim, one claim group will be generated; if multiple claims are generated, there are likely to be two claim groups.	Depending on how the care is logged in the EHR, there may be one encounter group or two (one for each day).

Findings

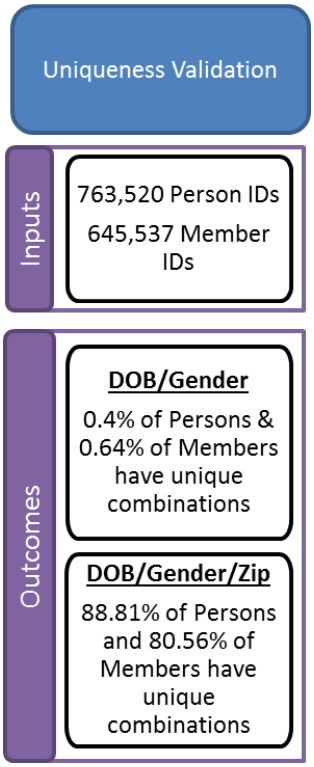
The grouping had the following impact:

- 16,066,402 claim lines were combined into 4,724,227 claim groups
- 3,760,625 encounter rows were combined 3,470,026 encounter groups

The grouping reduced the number of clinical events by a lesser degree than it reduced the number of claims events due to the fact that clinical data is pre-grouped into encounters during the load into the Arcadia Analytics data warehouse.

The team was initially concerned that grouping would have a distortive effect on the dates of service where a claim or an encounter spanned more than one day. For instance, in the event that an encounter lasted two days (as when a patient is admitted for an inpatient stay, then discharged the following day), and different claims were submitted for the first and second day of an encounter, the encounter group would carry a date of the first day only and could be matched to the first day's claim group, while the second day's claim group would go unmatched. However, only 292,389 encounters (7.78%) and 302,384 claim lines (1.88%) spanned multiple days; as such, the team determined that the impact of this distortion would be relatively low.

4.4. Uniqueness Analysis and Validation



To determine whether or not one could consider a member and a person a match based on their having the same demographic data, the team evaluated the uniqueness of the data. If, for instance, there were many persons in the clinical data with the same date of birth and gender, the team would know that, in trying to match a member to a person, using date of birth and gender would result in a high number of false positive matches (as is, in fact, determined to be the case).

Method

The team considered two possible combinations of demographic information:

- Date of Birth and Gender
- Date of Birth, Gender, and Zip Code

The team evaluated the uniqueness of each of these combinations to determine whether or not there was likely to be more than one person/member with this demographic set in either of the datasets.

Zip Code

The use of zip code for matching is complicated by the fact that one person, or even one event, may have different zip codes in the claims and clinical data. The claims data contains the zip code at which the patient resided at the time that the claim was filed. This field behaves differently in the clinical data: when the HIN MPI merges multiple patient records into person records, it selects the most recent information on a person. Thus, the clinical data presents the *most recent* demographic fields as the primary ones. As a result, if a person has moved since an event of care, the claim would reflect their zip code at the time of the event of care, but the clinical dataset would show the most recent zip code that has been reported to HIN.

Analysis of the MPI data showed that HIN persons have an average of 1.13 distinct zip codes on file; 88% have exactly one zip code. It is worth noting that, as providers are added onto the HIE, so are patients; therefore, it is possible that a person saw a provider in July 2011 who was not on the HIE at the time, then moved to a new zip code, then saw a provider in January 2013 who was on the HIE. In this case, the claim would have the person's old zip code and the HIN MPI would have only their new zip code.

Findings:

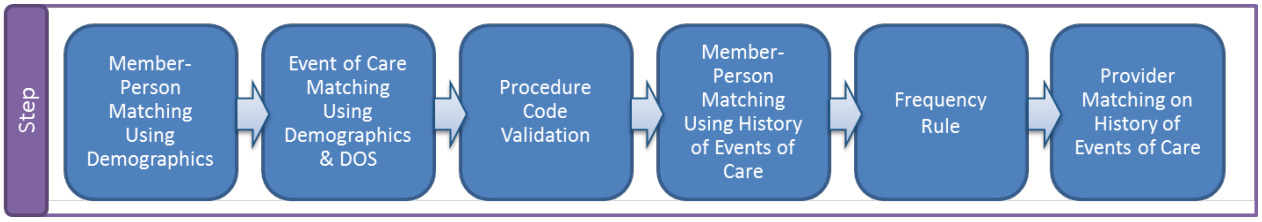
The team found that while date of birth and zip code repeat frequently in the population – a person has only a 0.4% or 0.64% likelihood (in clinical and claims data, respectively) of having a unique date of birth and gender. Zip code, on the other hand, adds significant differentiation: one's likelihood of there being no other person in the dataset with the same date of birth, gender, and zip is 88.9% or 80.6% (in clinical and claims data, respectively).

	Clinical Data	Claims Data
Total Number of Persons/Members	763,520 Persons	645,537 Members
Uniqueness of Date of Birth & Gender	There are 70,159 unique combinations of DOB and gender in the clinical data. 3,023 persons (only 0.40% of the population) have unique combinations of date of birth and gender.	There are 67,909 unique combinations of DOB and gender in the claims data. 4,108 members (only 0.64% of the population) have unique combinations of date of birth and gender.
Uniqueness of Date of Birth, Gender, and Zip Code	There are 718,603 unique combinations of DOB, gender, and zip code in the clinical data. 678,883 persons (88.91% of the population) have unique combinations of date of birth, gender, and zip code.	There are 579,568 unique combinations of DOB, gender, and zip code in the claims data. 520,021 members (80.56% of the population) have unique combinations of date of birth, gender, and zip code.

The fact that the persons in the clinical dataset are more unique than members in the claims dataset appears to confirm that two Member IDs in certain cases represent the same person, such as when a person has a primary and a secondary insurance plan, or when a person changes coverage and has claims under two different Member IDs.

Given that the HIE data suggests that changes in zip codes are not highly prevalent in the population, and that the uniqueness analysis demonstrates that not using zip codes will present a high instance of collisions within datasets and therefore false positives, the team proceeded to use zip code throughout the matching process.

5. Matching Methods and Findings

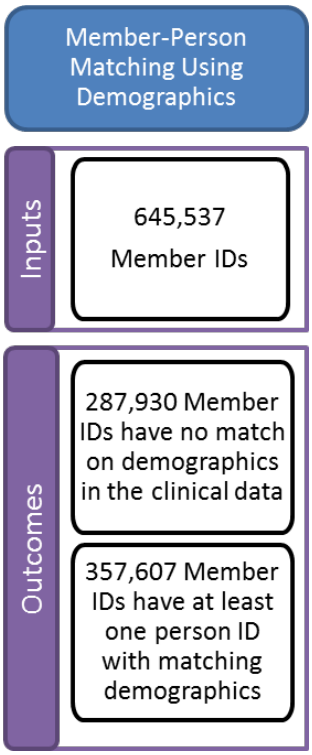


The team took the following steps to match the claims and clinical data:

- **Matching Claims Members to Clinical Persons** using demographics, eliminating those members who, judging by demographic information, were not present in the clinical data.
- **Matching Events of Care:** Identifying claim events and clinical events which, judging by the member or patient's demographic information and date of service, were present in the opposing dataset.
- **Procedure/Diagnosis Code Validation:** Confirming that matched event of care had largely consistent procedure and diagnosis codes.
- **Member/Person Matching Using Events of Care History:** Matching members and patients based on demographics and shared dates of service.
- **Frequency Rule Application:** In situations where demographics and dates of service matching resulted in multiple person matches for a single patient, identifying a single match or dismissing the member as not match-able. This is the final step in matching members and persons.
- **Matching Providers Using Events of Care History:** Using dates of service to match claims Provider IDs with clinical Provider IDs. This step establishes a provider ID crosswalk.

5.1. Member-Person Matching Using Demographics

The team first attempted to match HIN Person IDs and MHDO Member IDs based on demographic information.



Method

For this experiment, the team used a combination of demographic fields that included date of birth, gender, and zip code. A Member ID and a Person ID were considered a match if the values of these fields were equal.

One Person, Multiple Members

Due to the fact that the HIN Person IDs have been de-duplicated by the HIN MPI and the claims Member IDs have not, it is possible that a single HIN Person ID will match accurately to more than one claim Member ID. This is likely to be true in several cases:

- When a person has primary and secondary insurance policies;
- When a person switches insurance policies during the time period, as in a job change;
- When a person receives a new Member ID from their existing insurance policy, which may occur in this dataset as many plan years end on 12/31, allowing a person to switch plans effective 1/1; depending on the insurance carrier's internal processes, this may create a new Member ID.

For this reason, the team matched from the Member ID to the Person ID (searching the clinical data to find a match for the member in the claims data) for each member. There should be only one person and any matching activity that identifies two or more Person IDs matching a single Member ID indicates the presence of a false positive match. This could also indicate two true positives, or duplicate Persons that were not combined by HIN's MPI; for the purposes of the matching, the team assumes that there are no such cases.

Findings

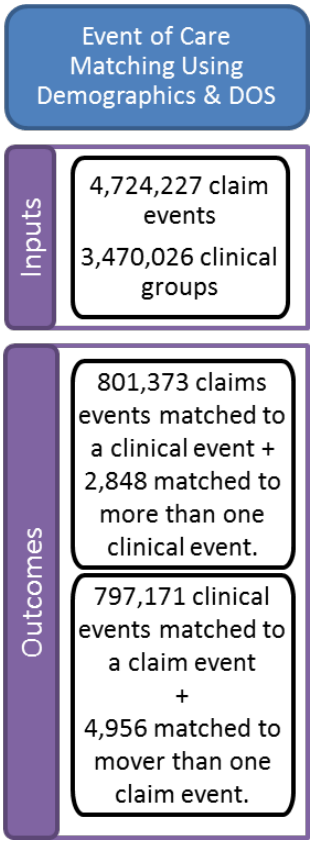
The team attempted to match 645,537 claims Member IDs to the 763,520 Person IDs in the HIE. Of these:

- 287,930 Member IDs (44.6%) have no match on demographics in the clinical data
- 357,607 Member IDs (55.4%) have at least one Person ID with matching demographics

This process has ruled out 44.6% of the Members as those who will not be matched to Persons.

5.2. Event of Care Matching Using Demographics and Date of Service

The next step was to match events of care represented by encounters on the clinical side, and claims on the claims side.



Method

For this experiment, the team used a combination of the demographic fields used above (date of birth, gender, and zip code) and date of service.

The risks associated with zip code discussed in Section 4.4 apply; in addition, the use of date of service is complicated by the impact of the grouping mechanism in situations where claim and encounter groups for the same event do not have a one-to-one relationship. For example, if an encounter spans two days but the procedures administered on those days are billed separately, it is likely that there would be a single encounter group (with the first day as the date of service) and two claim groups (one with the first day as the date of service, and another with the second day as the date of service). In this event, the encounter group would be matched with the first claim group, and the second claim group would not have a match.

Findings

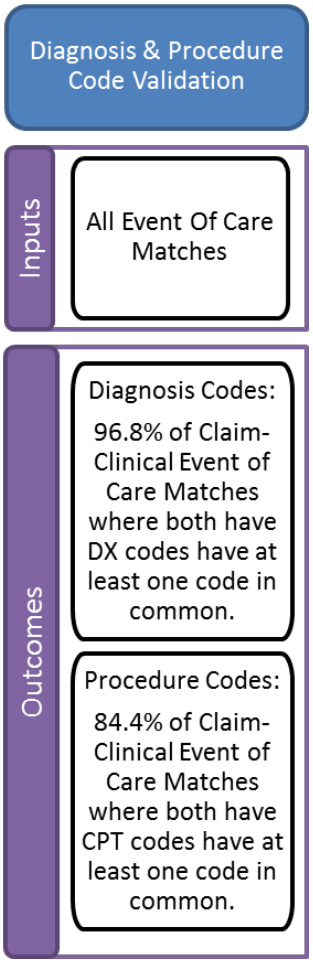
The results are as follows:

	Claims Data	Clinical Data
Number of Events of Care	4,724,227 claim groups	3,470,026 encounter groups
No Demographic + Date of Service Matching:	3,920,006 claims groups (83%) have no demographic/DOS match in clinical data	2,667,899 encounter groups (77%) have no demographic/DOS match in claims data
Successful Demographic + Date of Service Matching:	801,373 claims groups have a single demographic/DOS match in clinical data; 2,848 have more than one match.	797,171 encounter groups have a single demographic/DOS match in clinical data; 4,956 have more than one match.

Less than 1% of the claims or clinical events (0.06% of claims groups; 0.14% of encounter groups) find more than one match in the opposing dataset; this means that the Date of Birth, Gender, Zip, and Date of Service data points create a highly unique combination that is unlikely to generate false positive matches.

5.3. Diagnosis and Procedure Code Validation

Given the high uniqueness of the demographics/date of service combination, the team expected matches on these data elements to be true event matches with very few exceptions; as such, the team expected the same CPT and ICD-9 codes to be present on the matched claim and clinical events.



Method

The team compared the CPT and ICD-9 codes for claims and clinical events that had the same demographic information and the same date of service.

Findings

The findings indicate that two events of care happening on the same date, with the same date of birth, zip code, and gender, have an 85% likelihood of having the same procedure code and a 97% likelihood of the same diagnosis code.

	Procedure Code	Diagnosis Code
Matches containing at least one CPT/ICD9 Code:	289,353 matches contained at least one CPT code in both the claim and the clinical encounter	557,473 matches contained at least one DX code in both the claim and the clinical encounter
Of those, matches with at least one code in common:	245,444 (84.4%) of those matches have at least one matching procedure code	539,746 (96.8%) of those matches have at least one matching diagnosis code

One note of interest is that the Procedure/Diagnosis codes in the HIE are pulled from the EHR. Prior to the claim submission process, these codes are reviewed, and in many cases revised, by a billing specialist or agency to ensure that the claim is coded in adherence with the payer's policy and to maximize the payment to the provider; these revisions are particularly common with the procedure code.

It is possible that a more lenient comparison of code groupings – such as one where the CPT codes 99213 and 99214, which both indicate office or other outpatient visits for the evaluation and management of an established patient but rely on different levels of complexity of medical decision making, are considered equivalent – would yield even higher results without sacrificing accuracy. Data showed that family physicians choose 99213 for about 61% of visits with established Medicare patients and choose 99214 only about 23% of the time for the *same type*

of visit.[9] Accounting for the somewhat interchangeable nature of such groupings of codes may increase the accuracy of usage of procedure and diagnosis codes in validation. Additionally, use of data in provider billing systems may minimize the frequency of the NULL codes in clinical data and assist with this discrepancy.

Impact of Date of Service

A previous iteration of the matching experiments did not group encounters and claim lines into events of care in the manner described in Section 4.3. A key difference was that the earlier approach resulted in separate encounters for a single event of care; for example, an inpatient stay frequently presented as one encounter representing the admission, and another representing the discharge, each encounter with its own set of diagnosis and procedure codes.

This version of the matching experiment had significantly lower matches on diagnosis and procedure codes on encounters and claim lines matched based on demographics and date of service. One possible explanation is that diagnosis and procedure codes may change between the admission and the discharge. In the event of a patient presenting in the emergency room and being admitted, the admission codes may reflect an emergency room visit, while the discharge codes will reflect an inpatient visit. The team attempted to make the comparison less granular by truncating the codes such that, for example, the procedure codes representing emergency room visits (99282, 99283, 99284, and 99285) were reduced to the truncated code “9928”; even with this level of abstraction, the match rates were low.

The impact of this fragmentation was resolved by grouping the codes together; as the matching looked for the events to have at least one code in common, a claim event and a clinical were considered to match on procedure/diagnosis if at least one code was shared. In the event of an admission being coded as an ED visit and a discharge being coded as an inpatient admission, as long as the claim event reflected *either* an ED visit *or* an inpatient admission, the events are considered as matching on that procedure code.

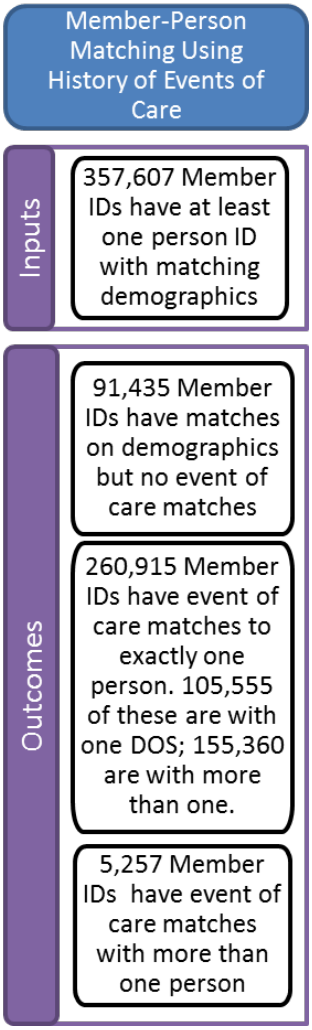
Conclusion

Given that a high number of matches did not contain CPT/ICD9 codes in both claims and clinical encounters, the team did not discount those event-of-care matches that matched on demographics and date of service but did not have any CPT/ICD9 codes in common.

[9] Jensen, Peter R. Coding "routine" office visits: 99213 or 99214? Before choosing 99213 for routine visits, consider whether your work qualifies for a 99214. Family Practice Management, v.12, no.8, 2005 Sept, p.52 (ISSN: 1069-5648)

5.4. Member-Person Matching On Event of Care History

The next step was to build upon the Event of Care matching to match members to persons.



Method

The team used the combination of demographics (Date of Birth, Gender, and Zip Code) and dates of service to match members to persons. Where a Member IDs combination of demographics and at least one date of service found only one Person ID with the same criteria, this Member ID was considered to match that Person ID. Given that the relationship between claim groups and encounter groups is not necessarily one-to-one, the team did not look for a complete match on date of service history; rather, a match of at least one date of service was considered a member/person match.

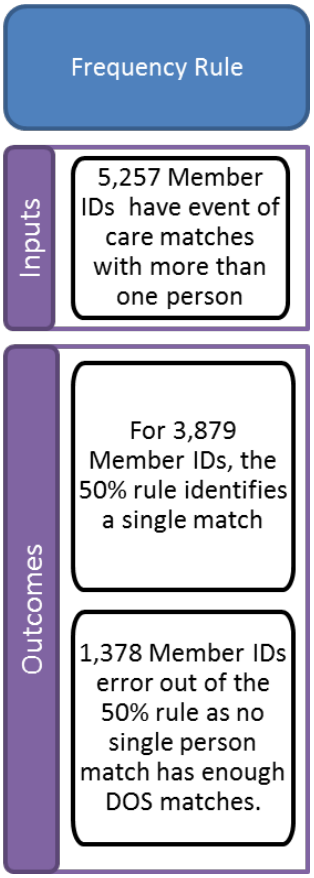
Findings

Of the 357,607 Member IDs for whom there were matches on demographics, 91,435 Member IDs (26%) were dismissed because the Person IDs with matching demographics did not have any events of care with the same dates of service; as such, while the demographics matched, there was no shared history to confirm the match.

The remaining Member IDs did find matches on demographics as well as date of service. Of these:

- 260,915 Member IDs (73%) saw only one Person match; there is no reason to believe that these matches are false positives.
- The remaining 5,257 Member IDs (1.47%), however, were matched to more than one person; for these, the team conducted a follow-up experiment to determine whether or not one of the person matches could be determined to be credible.

5.5. Frequency and the 50% Rule



Method

In the event that the demographics/date of service combination for a Member ID matched to more than one Person ID, the team applied a "frequency rule" to determine whether or not one of the matches could be considered a *better* match. The frequency rule states that, if a member is matched to multiple persons, only that person who matches on more than half of the matched events is considered a match. The following examples explain the approach, and the rationale:

Example 1:

Consider the following example below, in which a single member is matched to two persons; one Person ID has one event match, while the other has two event matches:

Claims Data					Clinical Data				
Member ID	**Date of Birth**	**Gender**	**Zip**	**Claim DOS**	**Person ID**	**Date of Birth**	**Gender**	**Zip**	**Encounter DOS**
M001	9/5/1974	F	04240	1/1/2012	P001	9/5/1974	F	04240	1/1/2012
M001	9/5/1974	F	04240	3/1/2012	P002	9/5/1974	F	04240	3/1/2012
M001	9/5/1974	F	04240	5/2/2012	P002	9/5/1974	F	04240	5/2/2012

In this instance, there are three event matches (as there are three claim events that are matched to clinical events); Person P001 has one event match (33% of the three event matches) and Person P002 has two event matches (66% of the three event matches). In this case, P002 is considered a *better* event match.

Example 2:

Consider the following example, in which a single member is matched to three persons, each on the basis of a single event:

Claims Data					Clinical Data				
Member ID	**Date of Birth**	**Gender**	**Zip**	**Claim DOS**	**Person ID**	**Date of Birth**	**Gender**	**Zip**	**Encounter DOS**
M001	9/5/1974	F	04240	1/1/2012	P001	9/5/1974	F	04240	1/1/2012
M001	9/5/1974	F	04240	3/1/2012	P002	9/5/1974	F	04240	3/1/2012
M001	9/5/1974	F	04240	5/1/2012	P003	9/5/1974	F	04240	3/1/2012

In this case, the total number of matches is three; each potential member-to-person match has one event match (33% of all event matches). In this case, no one match is better than the others, so no one match can be proven to be the match.

Example 3:

It is also possible that one match is better than another – but not better *enough.* Consider the following example, in which a single member is again matched to three persons:

Claims Data					Clinical Data				
Member ID	**Date of Birth**	**Gender**	**Zip**	**Claim DOS**	**Person ID**	**Date of Birth**	**Gender**	**Zip**	**Encounter DOS**
M001	9/5/1974	F	04240	1/1/2012	P001	9/5/1974	F	04240	1/1/2012
M001	9/5/1974	F	04240	2/1/2012	P001	9/5/1974	F	04240	2/1/2012
M001	9/5/1974	F	04240	3/1/2012	P001	9/5/1974	F	04240	3/1/2012
M001	9/5/1974	F	04240	4/1/2012	P002	9/5/1974	F	04240	4/1/2012
M001	9/5/1974	F	04240	5/1/2012	P002	9/5/1974	F	04240	5/1/2012
M001	9/5/1974	F	04240	6/1/2012	P003	9/5/1974	F	04240	6/1/2012
M001	9/5/1974	F	04240	7/1/2012	P003	9/5/1974	F	04240	7/1/2012

In this case, there are seven event matches in total. The match to P001 has three event matches (3/7 = 43% of all event matches), the match to P002 has two matches (2/7 = 29% of all event matches), and the match to P003 has two event matches (2/7 = 29% of all event matches).

In this case, if the selection of P001 as the better match (as it does, in fact, have more event matches than either of the other potential matches) would mean that more than half of this member's event matches (four event matches = 57%) were false positives. This implies that, for whatever reason, this member's data was unusually

common – too common to definitively match this member to a single person. **Hence the 50% rule requires that more than 50% of the matched events be matched to a single person.**

The 50% rule is somewhat arbitrary; it was selected by the team because, even if half the matches are attributed to a single person (say, one person had two event matches, and two other persons each had one event match), there would be as many false positive matches as true positives, and the team wanted to establish a rule which required there to be more true positive matches than false positives. It is possible, however, to execute the same matching activities with a higher standard by requiring a higher percentage of matches to be with a single person; alternatively, a lower percentage-of-matches requirement would match more Member IDs (with a higher risk of including false positives).

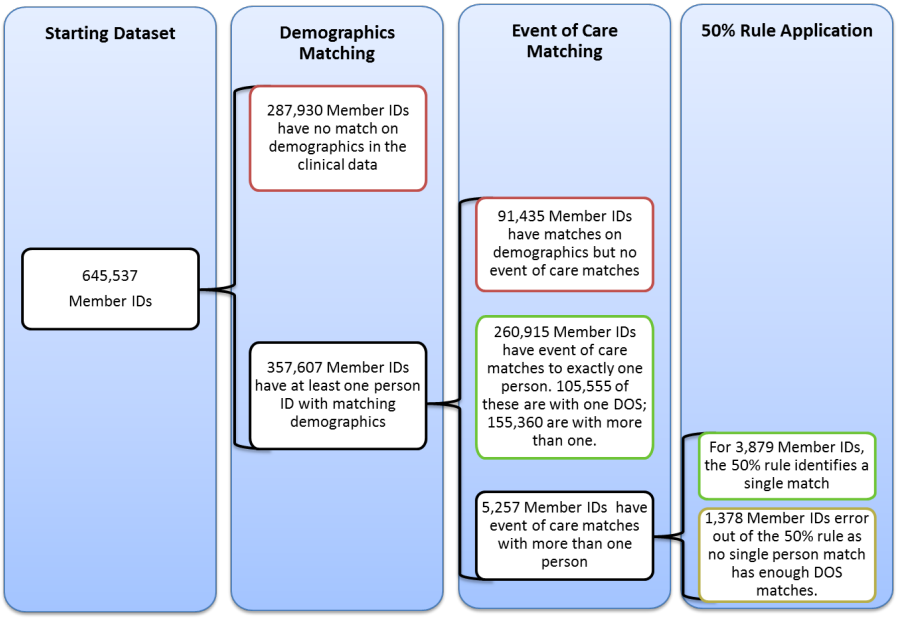
Findings

Of the 5,257 Member IDs that required the application of the 50% rule, 3,879 Member IDs were matched in such a way that more than 50% of all matched encounters of care were with a single Person ID. That is, for 3,879 Member IDs, the team used the 50% rule to dismiss all but one Person ID match, and considered that match to have sufficient frequency of events of care to consider it a strong enough match to overcome the doubt cast by the false positive matches.

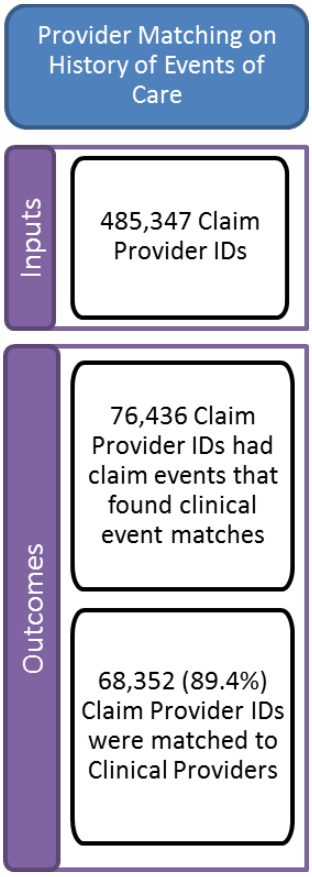
For the remaining 1,378 Member IDs, no one Patient ID has more than 50% of the matched events of care; for these, the matching process demonstrates that the matches are not sufficiently concentrated to demonstrate a single true match, and all matches for these Member IDs were dismissed as false positives. This may be indicative of duplicate persons in the HIN dataset: as the MPI uses more data elements than those that were available for this proof of concept, there may be persons who were not matched by HIN's MPI due to issues such as discrepancies in the spelling of the name. In this case, HIN may have two Person IDs for a single person with one Member ID. This may also indicate two individuals having the same demographic information (date of birth, gender, and zip code) and a shared date of service, which, while improbable, is not impossible.

Of the original 645,537 Member IDs:

- 379,365 Member IDs (55% of all) were not matched to a person due to lack of demographic + event of care matches
- Demographic and event of care matching found a match for 264,794 Member IDs (41%)
- 1,378 Member IDs (0.21%) error out, failing to find a single match that is strong enough to overcome false positives.



5.6. Matching Providers through Matched Events of Care



While the claims dataset does not have any identifying information about providers, event of care matches can be used to match providers by comparing dates of service of matching persons.

Provider Data

While the claims dataset does not identify providers directly, claims do have an encrypted provider field whose values are unique within each payer; therefore, if a single member has several encounters with a single provider, or if multiple members with the same payer see the same provider at the same facility, the encounters are labeled with the same Claim Provider ID, indicating that they are performed by a single provider.

As such, events of care matches can be used to match the Claim Provider IDs to the identified providers in the clinical data.

Method

For all claim events that were matched to a clinical event using Date of Service and Demographics, the Claim Provider IDs are associated to the Clinical Provider IDs on the associated clinical events.

In situations where a single Claim Provider ID is thus matched to multiple Clinical Provider IDs, the team again used frequency; in this case, the team used a 75% rule: If 75% of the matched events for a Claim Provider ID could be attributed to a single Clinical Provider ID, the team defined the two IDs as a match.

Findings

The claims data contained 485,347 Claim Provider IDs. Of these, only 76,436 had claims that found date of service match and demographic matches in claims data.

Of the 76,436 Claim Provider IDs that could be matched based on events of care, the team's methodology matched 68,352 (89.4%) Claim Provider IDs to Clinical Provider IDs. Of these, 95.3% could only be matched to a single Clinical Provider ID and did not require the invocation of a frequency rule. In the cases where the frequency rule was invoked, the Clinical Provider ID that was selected as the true match had an average of 7.76 event matches as compared to an average of 2.69 event matches on the Clinical Provider ID that was determined to be a false positive. The low occurrence of false positive matches in these findings, as well as the depth of matched event history, increased the team's degree of confidence in these results.

6. Conclusions

6.1. Findings

The overall results of the matching work were as follows:

Of 645,537 Member IDs, 264,794 IDs (41%) were matched to 254,120 Person IDs (33% of Person IDs).

380,743 Member IDs could not be matched. Of these:

- 287,930 had no matches on demographics in the clinical data, indicating that the member was not represented in the clinical data at all, or was represented with different demographic information (such as a more recent zip code or an inaccurately entered date of birth)
- 91,435 Member IDs did have demographic matches but no event of care matches. These may have been false-positive matches on demographics, or situations where a person had a claim that wasn't on the HIE and an encounter that wasn't on a claim. The latter is less likely, as the majority of events on the HIE that did not have associated claims were due to the payer's claims being unavailable; however, it is possible that an individual visited a primary care physician when they had Medicaid (generating a clinical encounter without a corresponding claim, as Medicaid data was missing from the claims dataset), then received insurance through Anthem and saw a chiropractor (generating a claim but no clinical encounter as chiropractic visits are underrepresented on the HIE). To measure the probability of this situation, one would need to understand the frequency with which individuals have encounters with *and* without commercial insurance with the timespan studied.
- 1,378 Member IDs (0.21%) were matched to multiple Person IDs based on demographics and date of service, and failed the 50% Frequency Rule, indicating too high a likelihood of either of the matches being a false positive to select a single match as the true match.

17% of claim events and 23% of clinical events were matched. Additionally, of 485,347 Claim Provider IDs, 68,352 were matched to Clinical Provider IDs.

While these volumes appear low, it is inherently limited by the constraints of the data. The claims dataset excluded approximately 48% of the population by the virtue of being limited to commercial insurance and, during the time period analyzed, HealthInfoNet covered an estimated 60-70% of care provided in Maine. The match rates confirm that the overlap in the encounters covered was limited, and each dataset benefited from the information contained in the opposing dataset.

6.2. Opportunities for Merged Dataset Leverage

The analysis confirmed that there is significant opportunity in further exploring the opportunities of linking the HIE data with the claims data held by the MHDO.

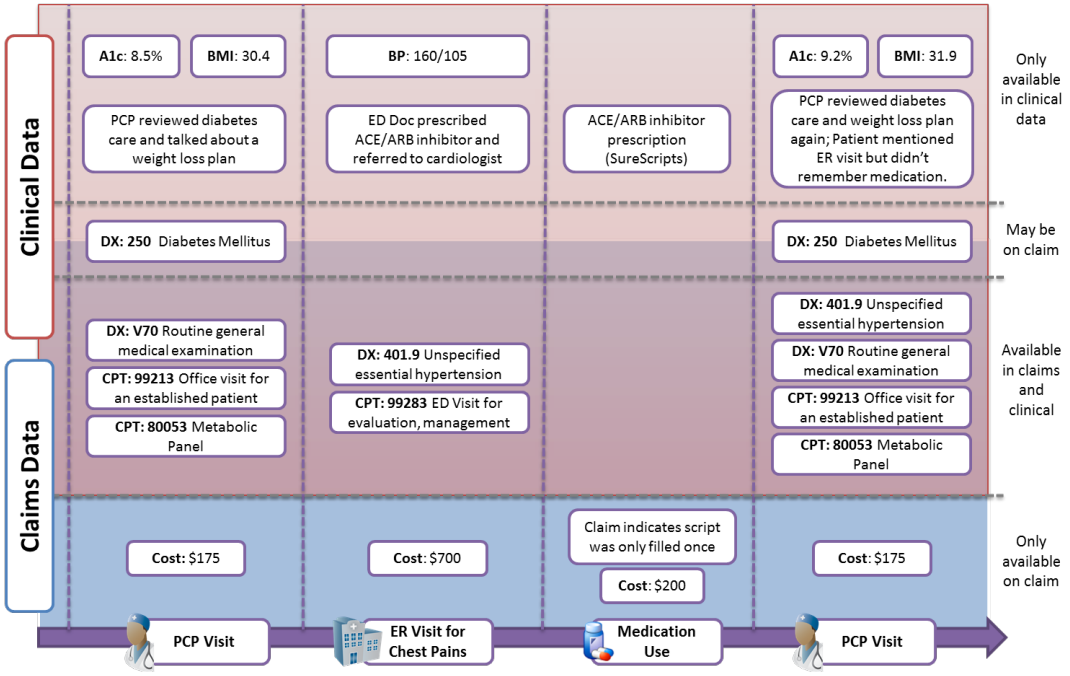
Broader Coverage: More Encounters

First, the sheer number of members/persons who do not have potential matches in the opposing dataset points to an opportunity to broaden the reach of either dataset by linking the two. As detailed in Section 3.3, there are events of care that are only present in one dataset or the other. Specifically, any analysis of claims data, even older datasets that contain Medicare and Medicaid data, excludes the uninsured population. Similarly, the HIN clinical dataset excludes out-of-state care, unconnected providers and persons that have opted-out of participation in the HIE (approximately 1%).

Deeper Insight into Care and Outcomes

For those encounters that are represented in both datasets, the linkage of the two sets offers a deeper view into each patient and each instance of care. The claims data presents the cost of individual encounters as well as total cost of care over the span of an episode or a lifetime, including costs to the health plan as well as the patient; this data is critical for driving cost containment efforts. In addition, the clinical dataset brings clinical outcomes such as vitals and lab results and often offers a more thorough accounting of an encounter than can often be found on the claim.

The following diagram demonstrates the claims and clinical views into a series of events for a hypothetical patient who experiences a series of fragmented events of care that are common when chronic disease management lacks coordination. This patient visits a primary care physician; the doctor takes note of his high A1C and BMI and advises him on weight loss and diabetes care. The patient may or may not follow the doctor's advice; some time later, he experiences chest pains and, fearing a heart attack, goes to the emergency room. This turns out to be a false alarm, but the ED physician diagnoses him with hypertension and prescribes an ACE/ARB inhibitor. The patient fills the prescription once, but never refills it. The patient goes on to visit his primary care physician, who may be unaware of the hypertension diagnosis or the prescription. In this story line, both the claims and clinical datasets individually paint a limited picture, but their compilation tells a fuller story:



This story may be further complicated by this patient switching insurances (in which case no one health plan would have a full claims history documenting his treatment) and by some of the providers involved not being on the HIE. Stories such as this one highlight the opportunities for use of claims and clinical data in care coordination: if these data sources are kept disparate, neither the health plan nor any of the providers realize that this patient was prescribed a medication meant to be taken on an ongoing basis that was only filled once. By overlooking this element of information, the patient's health plan and care team miss the opportunity to intervene and ensure that he follows his treatment plans appropriately.

Uses of Expanded Dataset

The merged dataset offers value for both research and care management. From a research perspective, the combination of cost of care information and clinical outcomes allows analysis on the cost and outcomes of care patterns – determining, for example, the impact of regular checkups on a diabetic's condition, or of participation in Patient Centered Medical Home on a patient's cost of care over time. In care management, the breadth of data allows a provider to see a patient's entire history of care, including events not on the HIE. In addition to encounters with out-of-state providers or those not participating in the HIE, claims data offers insight into medication compliance, a key aspect of patient engagement that directly impacts clinical outcomes and is currently nearly invisible to providers. Furthermore, as we introduce payment reform activities in the State, these data will become increasingly important to all parties to assess, validate and assure that value-based / accountable care purchasing activities have the desired impact – higher quality at a lower cost.

6.3. Furthering Matching Analysis

The following opportunities exist for building upon this analysis and improving match rates:

Validation

The greatest gap in the analysis presented in this report is the team's inability to validate the results. Confirming the accuracy of the match rates, even for a limited group of persons/encounters, would allow the team to determine which experiments, or which subsets of data, generate false positives and/or false negatives. Given a baseline, the team could fine-tune the matching process to eliminate false matches and identify new ways of going after matches that were not identified.

Additional Data Elements for Matching

The claims dataset used for the pilot was limited to date of birth, zip code, and date of service; introducing additional elements, such as the following, could increase the accuracy of matching:

- **Provider/Facility:** a true crosswalk of HIE providers to the Claim Provider IDs could be used to confirm matches with greater accuracy than the provider crosswalk the team created based on matched encounters, as the crosswalk is inherently limited by and dependent on the matching work done in earlier sections.

- **Additional Member Identifiers:** Any additional data elements that can be made available on the member will allow the matching methodology to be significantly more precise. While the most value would obviously come from truly unique identifiers such as social security number, any additional information, such as first name or street name, would be beneficial in this endeavor. A rule revision allowing MHDO to share additional member information may improve the precision of this analysis to the point where demographic matches are sufficiently indicative of an actual match that they can be used in care management.

- **ICD9/CPT Codes Billed:** Given the low overlap of diagnosis and procedure codes between matched claim and encounter groups, there is a high likelihood that the ICD9/CPT codes on a claim and a clinical encounter match for a single event of care. If this is the case, having the ICD9/CPT codes from the billing systems would allow the use of these for matching, as, for a single event of care, the codes entered by a provider's billing specialist should match those on the claim with the payers.

Excluding Events That Won't Have Matches

As discussed in Section 3, both the claims and clinical datasets contain information for which there are no corresponding encounters/claims in the opposing dataset. The matching process could be improved if the datasets were narrowed to only that information which had the potential to be matched:

- **Clinical Data:** Clinical data could be filtered by payer, if this information were available more consistently in the HIE or through provider eligibility rosters or member eligibility lists from payers. Being able to limit the clinical data to those encounters for which MHDO would have claims would limit the risk of claim-to-clinical matching generating false positives.

- **Claims Data:** Additional information on providers would be helpful in eliminating those claims whose providers are not on the HIE. If the providers were identified on the claims, the team would be able to filter the claims data by those providers who are a part of HIN. Alternatively, if this data were not made available by MHDO, HIN could provide a list of its member providers to MHDO, requesting that MHDO filter the claims data. Though less precise, provider specialty and provider address could also be used to filter out claims from providers who are less likely to be on the HIE. The exclusion of claims from providers not on the HIE would likely result in the ability to match a significantly higher portion of Claim Provider IDs to Clinical Provider IDs.

Exclude Situations Likely to Cause Collisions

There are specific situations that could be excluded from future analyses, as they are likely to generate false-positive matches:

- Persons and encounters with the date of birth of January 1, as this appears to be documented inaccurately for some individuals (excluding the Somali populations discussed above).
- Events within a short time period (such as a month) of a person's date of birth: while it is unusual for two people with the same birthday to have an event of care on the same day, it is less unusual that two infants born on the same day have an encounter on the day of their birth or the following day.

Cyclical Therapy Events

Cyclical therapy events may be handled differently from other events of care. Cyclical therapy events may not be recorded in the clinical data the same way as other events. Some facilities record multiple such events as a single encounter. For example, a person receiving dialysis on a weekly basis for several months may have a claim for each time they receive analysis, but the facility may check them in on the EHR, logging a new encounter, on the first of every month. In this case, the dates of service of claims and clinical events would differ.

If these events are recorded as discrete encounters, they still vary from other types of encounters in that a member and a person sharing multiple dates of service for a series of cyclical therapy events does not necessarily increase the likelihood that they are one individual. For example, it is typically more unlikely that two people with the same demographic set would have four events with the same date of service than that they would have one or two events with the same date of service; as such, a history of four identical events would typically indicate a higher probability of these being the same person. However, if two people receive weekly dialysis or radiation treatments on Wednesdays, once they've had an overlap of two events (both having received dialysis on a Wednesday, for example), the presence of a third or a fourth similar event (dialysis on the following Wednesday) does *not* increase the probability of these being the same person. As such, excluding series of regularly occurring events with the same procedure codes may reduce the probability of false matches that appear true due to a high number of matched events.

Greater Volume of Data

A longer timespan with claims and clinical data can also improve matching. The addition of another year of claims and clinical data would not significantly increase the number of Members/Persons, but it would increase the depth

of event of care history for the existing population. With a greater quantity of events of care, a greater quantity of matches is possible; while a person with only one claim has a relatively high probability that that one claim will not be on the HIE, when a person has many claims, it becomes less likely that none of their claims will be on the HIE. As such, applying the methodology to a longer time period is likely to improve the matching results.

Zip Codes

The matching may also be run with the original zip codes submitted to HIN with the encounter rather than those presented in the MPI. This will allow the team to use the zip code more accurately in matching, as the zip code in both the clinical and claims datasets would then reflect the patient's zip code at the time of the encounter.

6.4. Incorporating Identified Data

In addition to improving match rates using the methodology described above, organizations in Maine can move to achieve the outcomes of this proof of concept by merging identified claims dataset into the clinical data. For instance, HealthInfoNet can work with MHDO under contract to add specific identifiers for the claims data. While this will require addressing potential privacy concerns, it is the most direct way to explore the potential benefits of a merged claims/clinical dataset by simplifying the matching process and minimizing the risk of false-positive and false-negative matches.

Much of the methodology applied to these datasets is also likely to be applicable to the merging of two identified datasets in two ways:

- **Member/Person Matching:** Demographic matching may be required to match members and persons effectively if there are shortcomings in data thoroughness and quality. For example, the double-encrypted ID number was only present for ~70% of claims in this dataset; such gaps in data availability, as well as data quality issues such as inconsistent dates of birth on claims associated with a single Member ID, suggest that, even with a dataset with all data elements released and no encryption, some matching will still be necessary.
- **Event of Care Matching:** An identified dataset would also have the issue of claims and encounters not having a one-to-one relationship; a single claim may cover multiple encounters, and, likely more frequently, a single encounter may result in the generation of multiple claims. Grouping based on date of service within each dataset to generate distinct events of care and matching on dates can address this issue.

Appendix A: Data Frequency Analysis

Data Element: Date of Service					
Clinical			Claims		
Value	Records (Total & % of all encounters)		Value	Records (Total & % of all claim lines)	
NULL	164,278	4.37%	9/1/2011	85,785	0.50%
5/1/2012	18,233	0.48%	10/17/2011	80,700	0.47%
5/29/2012	17,034	0.45%	10/3/2011	80,690	0.47%
5/22/2012	16,816	0.45%	11/1/2011	80,669	0.47%
6/26/2012	16,601	0.44%	11/28/2011	79,541	0.46%
2/27/2012	16,595	0.44%	10/24/2011	79,335	0.46%
5/14/2012	16,578	0.44%	11/21/2011	79,123	0.46%
6/4/2012	16,550	0.44%	10/18/2011	79,064	0.46%
5/15/2012	16,508	0.44%	9/27/2011	79,048	0.46%
5/21/2012	16,501	0.44%	9/26/2011	78,592	0.46%

Data Element: Zip Code					
Clinical			Claims		
Value	Records (Total & % of all encounters)		Value	Records (Total & % of all claim lines)	
04240	48,221	1.28%	04401	486,371	2.82%
04401	42,195	1.12%	04103	461,822	2.68%
04330	37,176	0.99%	04240	440,094	2.55%
04901	32,633	0.87%	04106	395,626	2.29%
04210	30,953	0.82%	04330	335,069	1.94%
04005	28,688	0.76%	04074	309,466	1.79%
04103	27,170	0.72%	04210	305,796	1.77%
04011	26,467	0.70%	04901	300,875	1.74%
04106	22,339	0.59%	04011	295,185	1.71%
04072	22,313	0.59%	04072	288,437	1.67%
04240	48,221	1.28%	04401	486,371	2.82%

Data Element: Date of Birth					
Clinical			Claims		
Value	Records (Total & % of all encounters)		Value	Records (Total & % of all claim lines)	
NULL	2,322	0.0617%	8/27/1959	3,780	0.0219%
1/1/1987	203	0.0054%	1/12/1952	3,447	0.0200%
1/1/1950	177	0.0047%	6/7/1947	3,211	0.0186%
1/1/1985	175	0.0047%	12/28/1967	3,111	0.0180%
1/1/1980	173	0.0046%	1/8/1951	2,839	0.0165%
1/1/1989	173	0.0046%	2/22/1953	2,780	0.0161%
1/1/1976	166	0.0044%	9/28/1959	2,637	0.0153%
1/1/1984	164	0.0044%	7/15/1952	2,590	0.0150%
1/1/1990	164	0.0044%	3/16/1973	2,553	0.0148%
1/1/1901	164	0.0044%	4/4/1952	2,552	0.0148%

Data Element: Primary Diagnosis							
Clinical				Claims			
Value		Records (Total & % of all encounters)		Value		Records (Total & % of all claim lines)	
401.9	Unspecified essential hypertension	575,558	15.30%	V70.0	Routine general medical examination at a health care facility	517,548	3.00%
250	Diabetes mellitus	535,478	14.24%	V76.12	Other screening mammogram	319,057	1.85%
272.4	Other and unspecified hyperlipidemia	448,820	11.93%	250.00	Diabetes mellitus	294,457	1.71%
427.31	Atrial fibrillation	395,237	10.51%	739.1	Nonallopathic lesions, cervical region	288,090	1.67%
V76.12	Other screening mammogram	361,404	9.61%	V20.2	Routine infant or child health check	237,162	1.37%
V70.0	Routine general medical examination at a health care facility	361,028	9.60%	272.4	Other and unspecified hyperlipidemia	227,164	1.32%
599	Other disorders of urethra and urinary tract	313,640	8.34%	739.3	Nonallopathic lesions, lumbar region	212,810	1.23%
244.9	Unspecified hypothyroidism	225,709	6.00%	401.9	Unspecified essential hypertension	198,426	1.15%
V58.61	Long-term (current) use of anticoagulants	216,838	5.77%	724.2	Lumbago	187,136	1.08%
V20.2	Routine infant or child health check	194,176	5.16%	V04.81	Need for prophylactic vaccination and inoculation against influenza	172,007	1.00%

Data Element: CPT Code							
Clinical					Claims		
Value		Records (Total & % of all encounters)			Value		Records (Total & % of all encounters)	
99213	OFFICE OR OTHER OUTPATIENT VISIT FOR THE EVALUATION AND MANAGEMENT OF AN ESTABLISHED PATIENT	518,543	13.79%		NULL	N/A	1,152,156	6.68%
NULL	N/A	466,638	12.41%		99213	OFFICE OR OTHER OUTPATIENT VISIT FOR THE EVALUATION AND MANAGEMENT OF AN ESTABLISHED PATIENT	1,070,112	6.20%
99214	OFFICE OR OTHER OUTPATIENT VISIT FOR THE EVALUATION AND MANAGEMENT OF AN ESTABLISHED PATIENT	439,409	11.68%		99214	OFFICE OR OTHER OUTPATIENT VISIT FOR THE EVALUATION AND MANAGEMENT OF AN ESTABLISHED PATIENT	690,287	4.00%
36415	COLLECTION OF VENOUS BLOOD BY VENIPUNCTURE	369,178	9.82%		36415	COLLECTION OF VENOUS BLOOD BY VENIPUNCTURE	496,557	2.88%
99212	OFFICE OR OTHER OUTPATIENT VISIT FOR THE EVALUATION AND MANAGEMENT OF AN ESTABLISHED PATIENT	212,522	5.65%		97110	THERAPEUTIC PROCEDURE	395,475	2.29%
99283	EMERGENCY DEPARTMENT VISIT FOR THE EVALUATION AND MANAGEMENT OF A PATIENT	124,328	3.31%		98941	CHIROPRACTIC MANIPULATIVE TREATMENT	346,937	2.01%
85025	BLOOD COUNT	110,560	2.94%		90806	INDIVIDUAL PSYCHOTHERAPY	326,494	1.89%
80053	COMPREHENSIVE METABOLIC PANEL	75,134	2.00%		97140	MANUAL THERAPY TECHNIQUES	281,023	1.63%
99284	EMERGENCY DEPARTMENT VISIT FOR THE EVALUATION AND MANAGEMENT OF A PATIENT	72,252	1.92%		80053	COMPREHENSIVE METABOLIC PANEL	270,365	1.57%
85610	PROTHROMBIN TIME	70,309	1.87%		85025	BLOOD COUNT	256,781	1.49%